Kybernetika

Petr Volf

Analysis of generalized residuals in hazard regression models

Kybernetika, Vol. 32 (1996), No. 5, 501--510

Persistent URL: http://dml.cz/dmlcz/124816

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 $\boldsymbol{N}(t) = (N_1(t), \dots, N_n(t))'$ 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 $\lambda_i(t)\,\mathrm{d}t = P(\mathrm{d}\,N_i(t) = 1 \mid \sigma(t^-))$, 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 $i = 1, \dots, n$ 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, $i = 1, \dots, n$, 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 $(L_i(S_i), \delta_i)$, $i = 1, \dots, n$, 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}, $j = 1, \dots, N_i(\mathcal{T})$) 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 $L_i(t) - N_i(t)$, $t \in [0, \mathcal{T}]$ 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 $S \subset \{1, \dots, n\}$. Its intensity process is given by $\overline{\lambda_S}(t) = \sum_{i \in S} \lambda_i(t)$ and the time scale transformation $\overline{L_S}(t) = \int_0^t \overline{\lambda_S}(s)\, \mathrm{d}s$ 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 $0 \le T_{(1)} \le T_{(2)} \le \cdots \le T_{(m)} \le \mathcal{T}$ 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 $\widehat{\overline{H_S}}(t)$ (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 possible 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:

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

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 $d\widehat{\boldsymbol{B}}(s)$ 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 $s \in [0, \mathcal{T}]$ 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 $s \in [0, \mathcal{T}]$.

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.

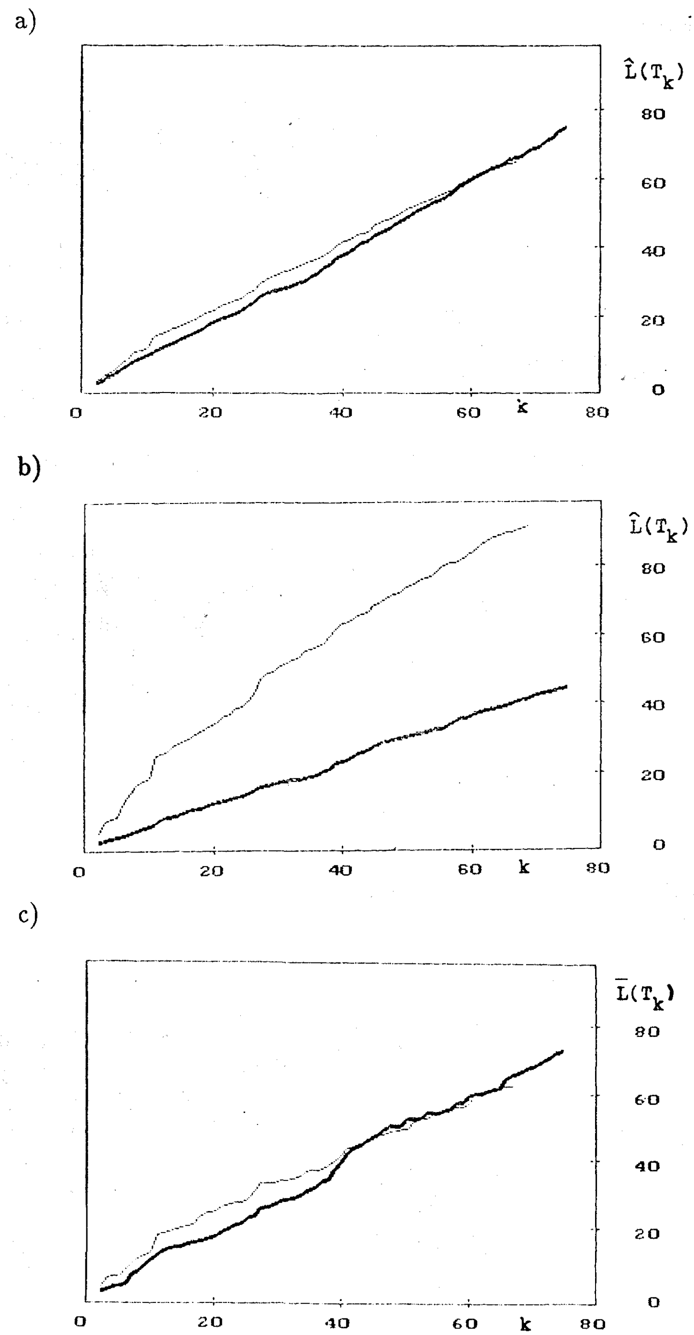

Fig. 1. Cumulative hazards: a) of the full model; b) of reduced model; c) of the 'true' model. Fine line – items with $X_2 < 10$, thick line – items with $X_2 \geq 10$.

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 $K(t) = C(t)/(1 + C(t))$, $t \in [0, \mathcal{T}]$. 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, 182 08 Praha 8. Czech Republic.