INFORMATICA, 1993, Vol.4, No.1–2, 111–125

PROPORTIONAL INTENSITY MODEL FOR REGRESSION IN EVENT–HISTORY ANALYSIS[1]

Petr VOLF

Institute of Information Theory and Automation
Czech Academy of Sciences
Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic

Abstract. The additive regression function is considered in the framework of the proportional hazard regression model for event-history data. The model is subjected to nonparametric estimation by the local likelihood procedure. Example illustrates the method, the hypotheses about the model are tested.

Key words: additive regression, counting process, event-history analysis, local likelihood (scoring), proportional intensity model.

1. Introduction. Let us consider a process of events occurring in time. The waiting time to the occurrence of a specified event is observed and examined. It is modeled as a random variable T, with continuous probability distribution function $F(t)$ and intensity – or hazard rate $h(t) = -d\log(1 - F(t))/dt$. The model for survival data from a biological research can serve as an example. However, the processes of waiting (or of duration) are encountered in many other areas, e.g., in demography, reliability engineering, economic and social surveys.

The interest of a statistical analyst is often concentrated to the estimation of how the distribution of T depends on other measured covariables. The specification of regression is frequently a part of the model for intensity. It means that an intensity function $h(t, x)$ is considered for the distribution of random variable T, when the

[1] The research was supported by the Czech Academy of Sciences grant 27 557

value of covariable equals x. Rather natural idea of separating the "baseline" common intensity from the influence of covariates gave rise to the proportional hazard model for regression. There, the semiparametric Cox's model is the most popular representant. It assumes that the intensity has the form $h(t,x) = h_0(t) \cdot \exp(\beta x)$.

However, from the moment when the Cox's regression model started its successful career in survival analysis (D.R. Cox, 1972), there were the attempts to treat the proportional hazard regression model more generally. This approach considers the nonparametric function $b(x)$ for the logarithm of hazard, the model is then given by the intensity

$$h(t,x) = h_0(t) \cdot \exp(b(x)), \tag{1}$$

where x is a value of covariate and t is a time to failure, for instance.

There have been derived methods and procedures determined especially for the models of survival, e.g., the kernel estimation of cumulative hazard function $H(t,x) = \int_0^t h(s,x)ds$ for a given x. The estimate is simply computed from results lying in a strata (the neighbourhood) around x. The results from this strata are considered as a homogeneous sample, they may be weighted by a kernel function. In Volf (1990 a) the method has been followed by procedures for estimation of component functions in model $H(t,x) = H_0(t) \cdot B(x)$, the cumulative version of (1). McKeague and Utikal (1991) have developed a serie of tests for discrimination between proportional hazard model and other specific forms of intensity model. Their tests are based on the results mentioned above, and on the concept of doubly cumulative hazard function. It means that $H(t,x)$ is secondarily integrated w.r. to x.

However, the global kernel estimation in moredimensional space of covariates has low effect. Even for $K = 3$ the realized points are rather sparse in R_K, great amount of data is needed in order to fill it sufficiently. That is why the idea of additive influence of covariates is often used. Some procedures specific for survival data model with additive logarithm of hazard proportion have been developed in Gentleman, Crowley (1991) and Volf (1990 b). We shall briefly recall the method. We shall then devote to the more

general concept modeling the event occurrence process by means of the counting process.

A survey of the theory and of statistical techniques for analysis of counting processes models is given in Andersen, Borgan (1985). Again, the development of a counting process is as a rule described by its intensity. In order to estimate it nonparametrically, we shall adapt and apply the local likelihood approach (Hastie, Tibshirani, 1986). Then the possibilities of testing are discussed for the specific form of the components of intensity. The method is illustrated by an example.

2. Estimation based on local likelihood. There exists quite a wide spectrum of various regression models. Nevertheless, the problems with their indentification are very similar. That is why there are attempts to deal with regression functions generally in order to obtain some general results, which then may be applied to specific cases.

Let the general regression function be some smooth function $b(\mathbf{x})$, describing the dependence of a response variable Y on a covariate $\mathbf{x}$. The additive regression model means that for $\mathbf{x} = (x_1, \ldots, x_K)^T$ with values from R_K the regression function is expressable as

$$b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j).$$

The component functions b_j are now the objects of estimation. The local likelihood method (described and modified to the local scoring algorithm in Hastie, Tibshirani 1986) consists in the following: If we wish to estimate the value $b_k(x_k)$ at the point $x_k = z$, we take function $b_k(\cdot)$ as a constant (b_z) in some chosen neighbourhood $\mathcal{O}_z$ around z. Let us suppose that we are able to construct the log-likelihood (ℓ_n) based on a random sample $\{Y_i, \mathbf{x}_i, i = 1, \ldots, n\}$. If b_z is treated as a parameter, we have to solve the equation $\partial \ell_n / \partial b_z = 0$ in order to estimate it.

In the most of common probabilistic models the logarithm of

likelihood can be expressed as

$$\ell_n \sum_{i=1}^{n} \ell_1(Y_i, b(\mathbf{x}_i)). \tag{2}$$

It follows that

$$\frac{\partial \ell_n}{\partial b_z} = \sum_{i=1}^{n} \mathbf{1}[x_{ki} \in \mathcal{O}_z] \cdot \ell_1'\left(Y_i, b_z + \sum_{j \neq k} b_j(x_{ji})\right).$$

It is seen that the local log-likelihood equation can be solved for b_z provided the estimates of remaining functions b_j, $j \neq k$, are available from previous step of estimation. Thus, the approach leads to an iterative algorithm, which starts from some initial guess about the functions, say $b_j^{(0)}(x_j) \equiv 0$ (or from $b_j^{(0)}(x_j) = \hat{\beta}_j x_j$, where $\hat{\beta}$ is the maximum likelihood solution for connected linear function).

As a rule, the equation is solved numerically by means of Newton–Raphson procedure, which needs the second derivatives of the likelihood. Schematically, the step from s-th to $(s+1)$-th iterative estimate can be expressed as

$$b_z^{(s+1)} = b_z^{(s)} - \frac{\partial \ell_n}{\partial b_z} \Big/ \frac{\partial^2 \ell_n}{db_z^2}. \tag{3}$$

Hastie and Tibshirani (1986) recommended for their local scoring to incporporate a smoothing directly into every step (3), to smooth also the derivatives of ℓ_n. Stone (1986) deals exclusively with the models of the form (2), especially with the exponential family of distributions. It assumes that $\ell_1(y, \theta) = Yc(\theta) + d(\theta)$, where c, d are known functions. Let $\theta = \theta(\mathbf{x})$ describe the dependence of Y on $\mathbf{x}$. Stone shows that under suitable conditions the following holds:

1. There exists the best additive approximation $\sum_{j=1}^{K} b_j(x_j) + b_0$ to function $\theta(\mathbf{x})$, with respect to the likelihood-based (Kullback-Leibler) distance.
2. This approximation can be estimated consistently by the polynomial splines.

Generally, the components b_j are ambiguous as to a shift. Stone considers norming conditions $Eb_j(X_j) = 0$ (i.e. covariates are the regarded as the realizations of random variables).

The local scoring algorithm naturally differs from global solution for reparametrized model, suggested by Stone. There is a hope that the solution is consistent, too, in the case of exponential family likelihood, although the proof is not given. The fact is well known at least for the Gaussian model, where the local likelihood coincides with the kernel estimation of regression function in traditional sense. Besides, if the "trivial" spline of order 0 (i.e. the histogram) approximation for functions $b_j(x_j)$ is considered, its global maximum likelihood equations are very close to the local likelihood ones. The difference is caused only by the use of fixed windows (in the case of histogram) instead of a moving window in the local likelihood approach. The difference vanishes asymptotically. Thus, the conviction about the consistency is well justified from this point of view. However, the numerical procedures of solution may differ considerably. The standard likelihood estimates of parameters are as a rule computed from one multivariate system of equations, meanwhile the local likelihood procedure solves recurrently a sequence of simple equations for one variable.

3. Proportional hazard regression model. As it was mentioned in Introduction, the model (1) is a popular (and natural) choice for description of influence of a covariate onto the hazard rate. Let us consider a frequently encountered design of survival data. A random sample $\{T_i,\ \delta_i,\ x_i,\ i = 1,\dots,n\}$ is observed, where T_i is observed value of time, x_i is a value of covariate and δ_i is the indicator of censoring from the right side. It means that $\delta_i = 1$ when T_i is a survival time, $\delta_i = 0$ if T_i is less than survival time, the i-th observation is censored at the time moment T_i. The inference for the hazard proportion $b(x)$ is based on logarithm of Cox's partial likelihood, namely on

$$\ell_n = \sum_{i=1}^{n} \delta_i \log \left\{ \frac{\exp b(x_i)}{\sum_{j=1}^{n} \exp b(x_j) \cdot I_j(i)} \right\}, \quad \text{where}$$

$$I_j(i) = 1 \text{ if } T_j \geqslant T_i, \qquad I_j(i) = 0 \text{ otherwise.}$$

It is seen that the likelihood is no more of the form (2). There were made several attempts to solve nonparametrically the estimation

task for this specific model. Let us mention two modifications of one method. They are based on the fact that the baseline cumulative hazard function $H_0(t) = \int_0^t h_0(s)\,ds$ is a part of the likelihood. Both procedures use two-step (alternating) iteration. One step computes estimate of H_0 provided function b has already been estimated. This step is common to both procedures, it utilizes a well known estimator

$$\hat{H}_0(t) = \sum_{i=1}^{n} \delta_i \cdot \mathbf{1}[t \leqslant T_i] \Big/ \sum_{j=1}^{n} \exp\left(b(x_j)\right) \cdot I_j(i). \tag{4}$$

The second steps differ formally. Gentleman and Crowley (1991) use the full log-likelihood (or this part of log-likelihood which is relevant), namely

$$\ell^* = \sum_{i=1}^{n} \delta_i \{b(x_i) + \log(h_0(T_i))\} - \sum_{i=1}^{n} \exp(b(x_i)) \cdot H_0(T_i).$$

The authors suggest a local solution $b(z)$ of $\partial \ell^* / db(z) = 0$, provided function H_0 has been estimated from the preceding step.

Volf (1990b) chose slightly different way. He noticed that random variables $U(x) = \ln \{H_0(T(x))\}$ fulfil the linear regression model

$$U = -b(x) + \varepsilon,$$

with ε distributed according to the standard doubly-exponential distribution. It again opens the way to the nonparametric (kernel-like, i.e. local) estimation of $b(z)$, from censored sample $\{V_i = \ln(\hat{H}_0(T_i)),\ x_i,\ \delta_i\},\ i = 1, \ldots, n$.

When the local maximum likelihood is used (w.r. to doubly-exponential distribution), the result of the method is quite identical with the result of Gentleman and Crowley. The procedure adapts easily to the case of the additive model.

However, this method is not adapted to the case in which the time-dependent covariates are included. The baseline C.H.F. $H_0(t)$ is no more a part of likelihood, function $b(x)$ has to be estimated from another source.

In the sequel, we shall consider a general design, based on the model of counting processes. Simultaneously, we enlarge the model and we allow the time-dependent (random) processes of covariates $\mathbf{X}_i(t)$, $i = 1, \ldots, n$. The counting process $N(t) = N_1(t), \ldots, N_n(t)$ is a set of right-continuous random step functions on $[0, T]$, 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 i-th kind or of i-th object in the life history. The upper bound T is such that $H_0(T) < \infty$.

The model is fully described by the (random) hazard rates for counting processes $N_i(t)$, namely $h_i(t) = h_0(t) \cdot \exp b(\mathbf{X}_i(t)) \cdot I_i(t)$, $i = 1, \ldots, n$, $t \in [0, T]$, where $I_i(t)$ is an indicator of the risk. The inference is again based on Cox's partial likelihood. Its logarithm is now

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

Let us again consider the K-dimensional covariate processes $\mathbf{X}_i(t) = X_{1i}(t), \ldots, X_{Ki}(t)$ and the additive regression function $b(\mathbf{x}) = \sum_{k=1}^{K} b_k(x_k)$. Let us fix the value z in the domain of, say, x_ℓ, and handle $b_\ell(x_\ell)$ like a parameter $b_\ell(z)$ in some neighbourhood $\mathcal{O}_z$ of z. Then the derivation of ℓ_n yields

$$\frac{\partial \ell_n}{\partial b_\ell(z)} = \sum_i \int_0^T \left\{ \mathbf{1}[X_{\ell i}(t) \in \mathcal{O}_z] - \exp(b_\ell(z)) \cdot \frac{R_\ell(z, b, t)}{S(b, t)} \right\} dN_i(t),$$

where $R_\ell(z, b, t) = \sum_{j=1}^{n} \mathbf{1}[X_{\ell j}(t) \in \mathcal{O}_z] \cdot \exp\left\{ \sum_{k=1}^{K} b_k(X_{kj}(t)) \cdot \mathbf{1}[k \neq \ell] \right\} I_j(t)$ and $S(b, t) = \sum_{j=1}^{n} \exp\left\{b(\mathbf{X}_j(t))\right\} \cdot I_j(t)$. By solving the equation $\partial \ell_n / \partial b_\ell(z) = 0$, we obtain the following iteration step:

$$b_\ell^{(s+1)}(z) = -\log \left[\sum_{i=1}^{n} \int_0^T \frac{R_\ell(z, b^{(s)}, t)}{S(b^{(s)}, t)} \, dN_i(t) \right.$$
$$\left. \Big/ \sum_{i=1}^{n} \int_0^T \mathbf{1}[X_{\ell i}(t) \in \mathcal{O}_z] \, dN_i(t) \right].$$

There $b^{(s)}$ is the estimate of function b (of all its components) obtained from previous, s-th step of iteration. It is seen that (at least) the values of $b_k^{(s)}$ at each observed $X_{kj}(T_i)$ are needed provided $I_j(T_i) = 1$. T_i denote now the times of observed counts. The first iteration step may start from $b_1^{(0)} = \ldots = b_k^{(0)} \equiv 0$ or from another convenient initial guess.

As the logarithm of partial likelihood does not correspond to the form (2), the conclusions of Stone (1986) does not hold for the proportional hazard regression model. The consistency has not been proven up to now. Nevertheless, the method has been checked by a number of examples, simulated as well as with real data. The results are encouraging. Sometimes the approximation of functions b_k by the regression splines is preferred (cf. Sleeper, Harrington, 1990). However, as it was pointed out, both approaches (local likelihood as well as the reparametrization by splines) meet with the same theoretical problems, although their computation procedures may differ one from the other.

4. Example. Let us now demonstrate the usefulness of the local likelihood method. The example has been constructed artificially, nevertheless it can represent a real situation.

Let us assume that the longtermed survey has been done in a company, in order to obtain an information about the dynamics of employment, especially about the departures of employees. Let us consider the data describing the result of this survey. Their structure is

$$\{T_i,\ \delta_i, X_{1i}, \ldots, X_{4i},\ i = 1, \ldots, n\}.$$

Response variable T covers last 10 years, it is measured in months from 0 to 120. Its value, T_i, denotes either the moment of departure of employee (i), or the moment of censoring, but in this case mostly $T_i = 120$. Here i is the number of an employee. The indicator variable $\delta = 1$ when the employee was fired, $\delta = 2$ when the individual left his job voluntarily (retired employees are included in this group), $\delta = 0$ for remaining or censored employees. The covariables have the following meaning: X_{1i} is the age of the individual in years at T_i, X_{2i} is the length of previous employment in the company, up

to the moment T_i. It is measured in years, too. X_{3i} characterizes the category of the job: 1 – researcher, 2 – specialist, 3 – administration, 4 – technical staff, 5 – unqualified assisting employees. $X_{4i} = 1$ for men, $= 2$ for women. Variable δ is thus considered as an indicator of two competing risks ($\delta = 1$ or $\delta = 2$), and of censoring ($\delta = 0$). Both X_1 and X_2 are measured (in years) at the moment of an event (or of censoring), both are changing during the time. It means that for example the value of covariate X_1 for i-th person at time t (months) is $X_{1i}(t) = \max\{0;\ X_{1i} - (T_i - t)/12\}$ (in years), similar connection holds for X_2. Yet the indicator of risk set has to be introduced. Let us define $I_i(t) = 1$ for $t \in [\max\{0;\ T_i - 12X_{2i}\},\ T_i]$ – the period during which the person has been with the staff of the company, $I_i(t) = 0$ otherwise.

Now, we are interested in the estimation of intensities for both events (considered separately, or considered together as one event – leaving the job.) The data are prepared for the analysis of covariate effects to the intensities in the framework of the proportional hazard regression model, namely the model allowing time-dependent covariates.

Results. The secondarily smoothed shapes of estimated functions b_1, b_2, b_3 are graphically displayed in Figure 1. After every step of their iterative estimation, the optimal (least squares) lines have been constructed from points $\{x_{ki}, \hat{b}_k(x_{ki}),\ i = 1, \ldots, n\}$. The changes of the parameters of these lines have served as an indication of convergence of our iteration procedure. The procedure started from $b_k \equiv 0$, $k = 1, \ldots, 4$. After the fifth iteration the changes of the slope parameters were less than 10^{-3}, we decided to stop the computation. The results of final linear approximation are displayed in Table 1. The analysis has been performed separately for both observed events, i.e. for $\delta = 1$ and $\delta = 2$. The table contains also the estimates of correlation of x_{ki} and $\hat{b}_k(x_{ki})$ and variance of residuas of $\hat{b}_k(x_{ki})$ from the line. The fourth covariable acquired two values only, its influence can fully be described by a linear function.

No norming conditions have been laid on the component functions.

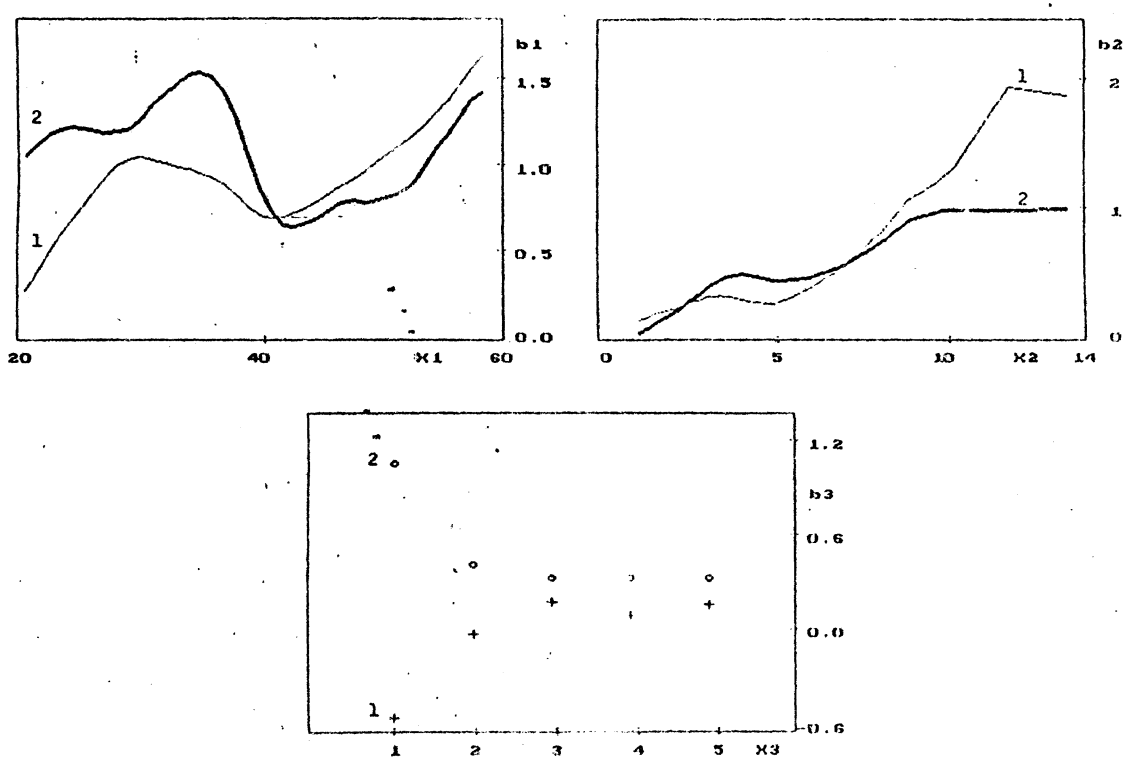

Fig. 1. Estimates of regression functions b_1, b_2, b_3 for risk $\delta = 1$, $\delta = 2$.

Table 1. Optimal lines approximating the functions $\hat{b}_k$.

	component (k)	intercept	slope	correlation	variance
$\delta = 1$:	1	0.3771	0.0137	0.5001	0.0549
	2	-0.1285	0.1218	0.8412	0.0435
	3	-0.5346	0.2004	0.7753	0.0251
	4	-1.0673	0.6258	1.0	0.0
$\delta = 2$:	1	1.6455	-0.0144	-0.3897	0.1129
	2	-0.2972	0.1131	0.8343	0.0396
	3	0.9571	-0.1991	-0.6960	0.0398
	4	0.6551	-0.4178	-1.0	0.0

The estimate of baseline hazard rate $h_0(t)$ in Figure 2 completes the graphical analysis. The cumulative version has been estimated directly from (4), then $\hat{h}_0(t)$ has been obtained by means of the kernel smoothing from $d\hat{H}_0(t)$.

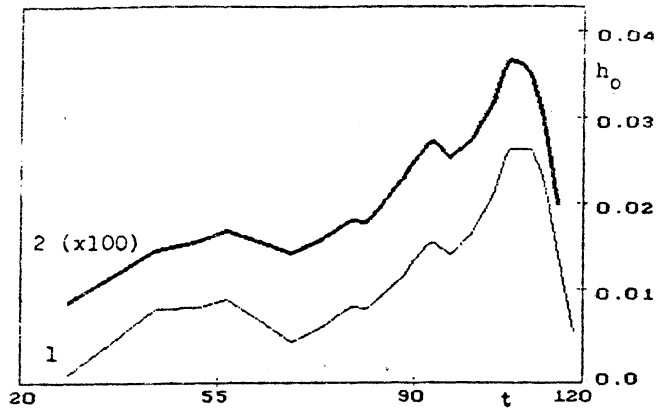

Fig. 2. Estimated baseline hazard rates h_0 for $\delta = 1$ and $\delta = 2$ (its scale has bee enlarged 100 times).

5. Concluding remarks about testing. Local likelihood (or local scoring, or moving window) method of estimation is able to reveal the shape of the general regression function. Provided the

proper model has been chosen. The correctness of the model can be tested on several levels. Let us demonstrate it on our example.

First, the correctness of the proportional 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 stratas around some values may be considered. Only for the sake of simplicity, let us test the assumption about proportional hazard dependence on the fourth covariate, X_4, which is only two-valued. The cumulative hazard functions have been estimated using the Nelson-Aalen estimator (separately for men and for women). The estimator is the special case of (4), when the sample is regarded as homogeneous, i.e., $b(x) \equiv 0$ is inserted into (4). Figure 3 shows the results. It is possible to admit that the two curves have approximately constant difference.

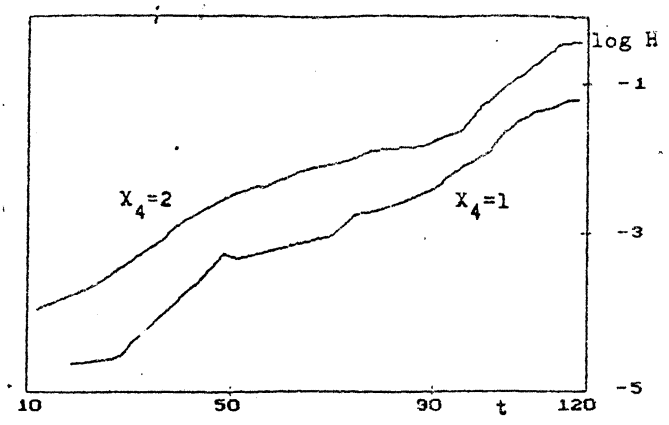

Fig. 3. Comparison of logarithms of cumulative hazard rates for men (1) and women (2).

Second, the significance of regression can be doubted. The hypothesis is tested that some regression functions b_ℓ are constant, that the dependence of response on corresponding covariates x_ℓ is negligible.

In the framework of Cox's model, the estimates of Cox's parameters β are asymptotically normal (cf. Andersen, Gill, 1982). 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 for a while that the Cox's model is the right one for our data. Table 2 displays the partial likelihood-based estimates of parameters β_k, $k = 1,\ldots,4$, together with the values of the test statistics G_k. The value of G_k should approximately come from the standard normal distribution provided the hypothesis $\beta_k = 0$ holds.

Table 2. Estimated Cox's model parameters and values of test statistics

	$\delta = 1$		$\delta = 2$	
k	β_k	G_k	β_k	G_k
1	0.0075	0.5726	-0.0259	-1.2259
2	0.0712	1.6528	0.1428	2.0139
3	0.1124	0.8628	-0.4773	-1.8863
4	0.7474	2.8004	-0.2057	-0.4932

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 $2\alpha = 0.1$, then $q(0.05) = 1.645$ is used as an approximate bound of critical interval. The results from Table 2 suggest the following conclusions: For the first event ($\delta = 1$), the hypothesis (that the risk does not depend significantly on a covariate) is rejected for components X_2 and X_4. When the risk of the second event (i.e., for $\delta = 2$) is considered, the hypothesis of negligible dependence is rejected for components X_2 and X_3.

With some license, the same conclusion 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 polynom. Then let us test the hypothesis that the coefficients of order higher than one are zero.

Another way of testing can be based on the linear approximation analysis of nonparametrically estimated regression function. Table 1 contains the results of such an analysis. However, the results depend strongly on the "smoothing policy" during the local likelihood iterations.

REFERENCES

Andersen, P.K., and O. Borgan (1985). Counting process model for life history data: a review (with discussion). *Scand J. Statist.*, **12**. 97–158.

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

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

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

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

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.*, **14**, 590–606.

Volf, P. (1990a). A large sample study of nonparametric proportional hazard regression model. *Kybernetika*, **26**, 404–415.

Volf, P. (1990b). Estimation procedures for nonparametric regression models of lifetime. In *Transactions of 11th Prague Conf.*, Academia, Prague. Vol.B. pp. 447–452.

Received December 1992

P. Volf graduated in mathematical statistics at the Charles University, Prague. In 1980 he received his Ph.D. in theoretical cybernetics. Later his interest has turned to the statistical methodology for regression analysis, with emphasis on lifetime regression models and nonparametric techniques. He is engaged both in theoretical and in applied research. at present he is a Research Scientist at the Institute of Information Theory and Automation, Prague.