

Regression splines for threshold selection in survival data analysis

Nicolas Molinari^{1,*}, Jean-Pierre Daurès¹ and Jean-François Durand²

¹ *Laboratoire de Biostatistique, Institut Universitaire de Recherche Clinique, 641, avenue Gaston Giraud, 34093 Montpellier, France*

² *Laboratoire de Biométrie, ENSA-M INRA UMII, 2 place Viala, 34060 Montpellier, France*

SUMMARY

The Cox proportional hazards model restricts the hazard ratio to be linear in the covariates. A survival model based on data from a clinical trial is developed using spline functions with variable knots to estimate the log hazard function. Moreover, the main point of the method is that a knot, seen as free parameters for a piecewise linear spline, represents a break point in the log hazard function which may be interpreted as a threshold value. The likelihood ratio test is used to select the final model and to determine the threshold number for a covariate. Confidence intervals for these threshold values are computed by bootstrapping the data. Two examples illustrate the method. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

In medical and industrial statistics, methods for evaluating the dependence of survival time or response time T on independent variable or covariate x have received considerable attention. The independent variable could be state of disease, duration of symptoms prior to treatment, or binary variable representing control or treatment group. The Cox model [1] is a popular choice for the analysis of censored survival data because it is semi-parametric, conceptually appealing, and efficient against proportional hazard alternatives. The modelled response is the hazard rate of failure, with a log hazard ratio linear in the covariate. The Cox regression model assumes that $\lambda(t, x)$, the hazard function of the continuous random variable t , is given by

$$\lambda(t, x) = \lambda_0(t)e^{x\beta}$$

where β is unknown parameter reflecting the effect of x on survival and $\lambda_0(t)$ is an unspecified arbitrary non-negative function of the time.

* Correspondence to: Nicolas Molinari, Laboratoire de Biostatistique, Institut Universitaire de Recherche Clinique, 641, avenue Gaston Giraud, 34093 Montpellier, France

However, this assumption is violated when covariate effects are best represented by smooth non-linear functions. A proportional hazard model incorporating an arbitrary covariate effect is of the form

$$\lambda(t, x) = \lambda_0(t)e^{g(x)}$$

where g , the log hazard ratio function, is an unspecified smooth function of x . Sleeper and Harrington [2] approximated g by a spline function expressed as a linear combination of B -spline basis functions with fixed knots. O'Sullivan [3] uses smoothing splines to estimate non-linear covariate effects in the Cox model. Smoothing splines are effective for examining fine details of regression data in exploratory analyses. In large data sets, methods using smoothing splines in likelihood-based regression models can require considerable computation resources, especially when cross-validation is used for choosing the smoothing parameter. Kooperberg *et al.* [4] use linear splines and their tensor products to estimate the log hazard function based on one or more covariates. The first three knots are placed at the quartiles of the uncensored data, and during the stepwise addition stage, new knots are successively added.

We use a B -spline representation for modelling the quantitative covariate effect, but knots are now seen as free variables to improve the fit. Moreover, to classify patients in groups, a physician usually uses his experience to determine threshold on variable values. The proposed method for survival data is based on the log hazard ratio function estimation with piecewise linear splines (degree one) with free knots. In fact, knot locations which correspond to a break point in the linearity indicate a change in slope for the log hazard ratio function. Then a knot location can be seen as a threshold value which corresponds, from a clinical point of view, to a change on the risk function. Because the linear model is nested in spline models, a likelihood ratio test allows selection of a model which corresponds to determine the threshold number. Moreover, a confidence interval for threshold values is computed by bootstrap resampling.

We illustrate the method with the Stanford heart transplant data presented by Miller and Halpern [5], which have been subsequently reanalysed by Hastie and Tibshirani [6] and by Durrleman and Simon [7]. The threshold determination is also presented on real lung cancer survival data.

2. THE PROBLEM

2.1. Stanford heart transplant data

Miller and Halpern [5] provided a number of analyses of the Stanford heart transplant data. The Stanford heart transplantation programme began in October 1967. By February 1980, 157 patients had received heart transplants. Of these 157 patients, 55 were still alive, that is, were censored as of February 1980, and 102 were deceased, that is, uncensored.

Patients are accepted into the programme when judged by physicians to be suitable candidates for transplantation. When a donor heart becomes available, medical judgement is used to select the patient who should receive it. The data consist of 157 observations of the final vital status indicator (alive or dead), the time to failure (months) and two covariates, age (years) and T5 mismatch score. Here we will only consider the age variable.

Various methods are compared with regard to their ability to fit the relation between age and survival in a proportional hazards model. Moreover, we determine a threshold value on the age variable.

2.2. Lung cancer data

Treatment of small-cell lung cancer (SCLC) is probably one of the great challenges of medical oncology owing to an increasing incidence in both men and women and a poor prognosis despite chemosensitivity. Serum markers have been proposed as a help in the management of SCLC during chemotherapy. In this setting, the most established serum marker is the gamma-gamma isomer of a glycolytic enzyme referred to as neuron specific enolase (NSE). High serum NSE and advanced tumour stage are well-known negative prognostic determinants of SCLC when observed at presentation. Recently, a tumour marker detecting cytokeratins in the serum was proposed [8]: CYFRA.

The relationship between risk of death and marker level during treatment of SCLC chemotherapy is not known. A total of 124 patients with SCLC were followed during cisplatin-based chemotherapy. The sample is (x_i, t_i, c_i) , $i = 1, \dots, 124$, where $x_i = (\text{cyfra}_i, \text{nse}_i)$ are two predictive variable values, t_i the survival time and c_i the final vital status indicator for the i th patient. Threshold values are computed that allow classification of the patients according to their marker values.

3. THE MODEL

3.1. Splines in a nugget

Using splines in a simple or multiple regressive model allows the investigation of non-linear effects with continuous covariates. In fact a spline function belongs to a finite dimensional linear space. To choose a basis for this linear space, B -spline basis functions are very appropriate due to the fact that they are numerically well conditioned and also because they achieve a local sensitivity to data.

Let $(\xi_0 =) a < \xi_1 < \xi_2 < \dots < \xi_K < b (= \xi_{K+1})$ be a subdivision by K distinct points on the interval $[a, b]$ on which the x variable is valued. These points are called the 'knots' of the spline function $s(x)$ used to transform the x variable. A spline is a polynomial of degree d (or order $d + 1$) on any interval $[\xi_{i-1}, \xi_i]$, and has $d - 1$ continuous derivatives on the open interval $]a, b[$.

For a fixed sequence of knots $\xi = (\xi_1, \xi_2, \dots, \xi_K)'$, the set of such splines is a linear space of functions with $K + d + 1$ free parameters [9]. A useful basis $\{B_l(\cdot, \xi)\}_{l=1, \dots, K+d+1}$, for this linear space is given by Schoenberg's B -splines, or *Basic*-splines [10].

A linear combination of B -splines gives a smooth curve. De Boor [9] proposed a recursive algorithm to compute B -splines of any degree from B -splines of lower degree. An example of basis elements with $d = 1$ and $\xi = (1, 2)$ is given in Figure 1.

We can now write a spline as

$$s(x, \beta, \xi) = \sum_{l=1}^{K+d+1} \beta_l B_l(x, \xi)$$

where the vector $\beta = (\beta_1, \dots, \beta_{K+d+1})'$ of spline coefficients is to be estimated from the data.

B -splines should rather be denoted by $B_l(\cdot, d, \xi)$, because they depend on both d and ξ , the vector of knots, which is considered as a tuning parameter. In practice, few well-located knots generally suffice in most cases, but deciding on their optimal location is a difficult problem due to the presence of local optima. However, when few knots are needed, which is the case for selecting a threshold, the problem of local optima is solved by successively processing some algorithms of

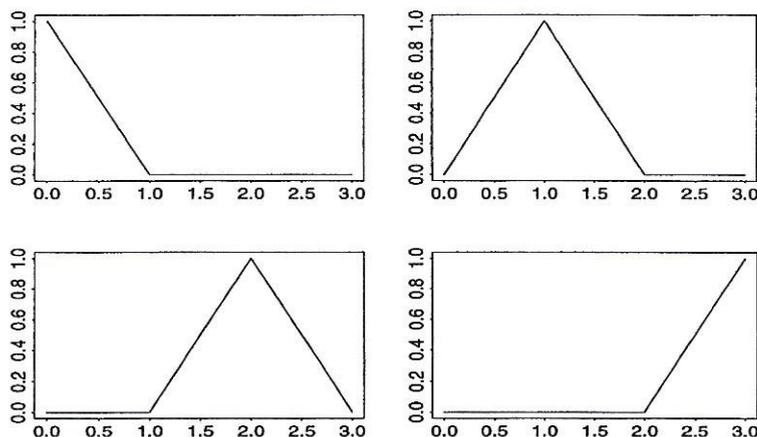


Figure 1. The B -spline basis on $[0, 3]$ with $d = 1$ and $\xi = (1, 2)$.

optimization initialized with different values. The section below details this approach adapted to Cox spline modelling.

3.2. The Cox regression model

For survival data analysis, the Cox regression model is a useful method, so it seems natural to use it to detect a threshold in quantitative variables with survival time. Moreover, using a spline function to detect non-linear effects needs to adapt the partial likelihood function.

We assume that data from a clinical trial of n patients $(x_1, t_1, c_1) \cdots (x_n, t_n, c_n)$, are at disposal. Thereby t_i denotes the observed survival time, that is, the interval for which the i th patient has been observed from entering the study until leaving. The binary status variable c_i indicates whether leaving has been through failure, like death, relapse or infection, or through censoring. Denote x the covariable, and x_i the covariate value for the i th patient. The Cox regression model assumes that the hazard of failure at time t for the i th patient is

$$\lambda(t, x_i) = \lambda_0(t) e^{x_i \beta} \quad (1)$$

The unknown parameter β is estimated by partial likelihood, that is, no further assumptions about the unknown baseline hazard function $\lambda_0(t)$ are imposed.

Usually, survival data of the type envisaged here are subject to right censoring because some individuals have not failed on termination of the study. It will be assumed throughout the paper that the censoring and failure mechanisms are independent. For the n individuals in the study with independent variable x_i , $i = 1, \dots, n$, let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ denote the ordered uncensored failure times with corresponding values $x_{(1)}, x_{(2)}, \dots, x_{(k)}$ and denote by $R(t_{(i)})$ the collection of individuals with censored or uncensored failure times $\geq t_{(i)}$.

Following Cox [1], we are interested in finding the estimators which maximize the partial likelihood

$$L(\beta) = \prod_{i=1}^k \frac{\exp(x_{(i)} \beta)}{\sum_{j \in R(t_{(i)})} \exp(x_j \beta)} \quad (2)$$

It follows that the log hazard ratio function (LHR) with respect to x is a linear function of x :

$$\text{LHR}(x) = \log \left(\frac{\lambda(t, x)}{\lambda_0(t)} \right) = x\beta$$

This supposes that a unit change in x has the same effect on the patient's log hazard ratio all over the range of x . This type of modelling is restrictive since the behaviour of $\text{LHR}(x)$ may be non-linear. The spline utilization provides more flexibility to model a continuous covariate.

3.3. Cox spline regression model with free knots

Unlike (1), to obtain more flexibility in modelling a continuous covariate, the Cox B -spline regression model is defined by

$$\lambda(t, x, \xi) = \lambda_0(t)e^{g(x)} \quad (3)$$

and the log hazard ratio function can be written

$$\text{LHR}(x) = \log \left(\frac{\lambda(t, x, \xi)}{\lambda_0(t)} \right) = g(x)$$

We approximate $g(x)$ by a spline function $s(x, \beta, \xi)$, and the log hazard ratio function is defined by

$$\text{LHR}(x) = s(x, \beta, \xi)$$

where the coefficients β and ξ are estimated by

$$(\hat{\beta}, \hat{\xi}) = \arg \max_{\beta, \xi} \prod_{i=1}^k \frac{\exp(s(x_i, \beta, \xi))}{\sum_{j \in R(t_i)} \exp(s(x_j, \beta, \xi))} \quad (4)$$

The spline partial likelihood function was computed using the S-Plus[®] language [11]. To solve (4), we use a classical maximizing gradient based method. Starting with $\beta = (\beta_1, \dots, \beta_{K+2}) = (0, \dots, 0)$ and with equally spaced knots ξ correspond to a null log hazard ratio function. To avoid the problem of local optima, different sets of initial knot values are located on a grid constructed within the range of the variable. We heuristically divide the range of the variable in 10. Because knots localized at the bounds of the interval are not an influence on the estimation, we are only interested in the nine inner values. Thus, we obtain $\binom{9}{K} = \frac{9!}{K!(9-K)!}$ different vectors of knots which are used for initializing the algorithm. We restart the minimization using these parameters and the null β vector. Note that initial ξ vectors with two or more confounded knots are not used because this choice allows us to obtain only a local minimum due to the lethargy theorem [12].

3.4. Thresholds determination

The preceding method is used to estimate the log hazard ratio function with linear splines. For splines of degree one, knots are points where the slope is changing in the shape of the piecewise linear function. These variations of the log hazard ratio function estimation are full of meaning. For example, if the estimated spline is constant down to the knot then increases quickly, it can be interpreted as the point separating the variable range in two parts. Patients with an x value lower than the knot location have a lower risk than the other, and the knot position ξ can be interpreted

as the threshold value. In practice, a low number of threshold values is of interest in medicine. In fact, only a model with one or two threshold values provides interpretable information because, generally, it is sufficient to classify patients in two or three groups.

It should be noted that spline models with K knots are nested in spline models having these K knots completed with some more different knots. However, using optimal knot locations obliged us to consider that the models are not nested. The number of degrees of freedom for free knot spline functions is a much debated question. In his paper, Owen [13] presents a summary on this subject. According to Feder [14], we define the degree of freedom of the spline model as $2K+d$ (K knots and $K+d$ for β coefficients). Note again that only a few number of knots is interesting because one more knot increases by two the degree of freedom. The fact that the classical linear model is nested in each spline model (see proposition 1 in Appendix), however, allows comparison of spline models ($2K+d$ degrees of freedom) to the linear model (1 degree of freedom). The classical likelihood ratio test gives us a p -value for each model against the linear. We select spline models with a significant p -value lower than 0.05. Otherwise the regression with a linear function is proposed to estimate the log hazard ratio function and a threshold value is not available.

Suppose the method is used with several linear spline functions with different number of knots, and two or more spline models are selected with a significant p -value, the problem is to find the optimal number which corresponds to the threshold number. A simple approach is to select the model with the more significant p -value. However, an effective method to make a comparison between two selected spline models with a different number of knots is to estimate for each model the distribution of the p -value. We perform again the Cox B -spline piecewise regression on 1000 bootstrapped samples and note the corresponding p -value. In fact, the 950th sorted value of the 1000 p -values provides an estimation of the 95th percentile of the p -value distribution. The model corresponding to the lower 950th sorted value is then selected. Note that this procedure also allows us to obtain a confidence interval for knots position. By noting on the 1000 bootstrapped samples the position of knots as well as the corresponding p -value, we estimate the knot distribution. The interval which corresponds to the 25th and the 975th of the 1000 sorted knot values represents a confidence interval at 95 per cent for the knots.

4. APPLICATIONS

4.1. Stanford heart transplant data

Table I summarizes the results of the various fitting procedures. Note that results for the local likelihood and local scoring methods appear in reference [6]. Results for the cubic spline approach are due to Durrleman and Simon [7].

From the goodness-of-fit point of view, Table I shows that our spline models provide optimal values very close to those from their competitive methods. Note, however, that the aim of our approach is to propose an effective method to detect accurately the threshold for heart transplantation. Thus we have to compare between Cox spline regression of degree 1 with one and two knots. Bootstrap results presented in Table II allow choice of the model with only one knot according to the p -value distribution.

Figure 2 shows the log hazard ratio function estimation for the selected model. Knot position (≈ 47 years) for the data is a break point of the log hazard ratio. Before the break point, the function seems roughly constant, and increases sharply. Thus we assume the age of 47 years is a

Table I. Analysis of Stanford heart transplant data.

Model	-2 log-likelihood	d.f.	<i>p</i> -value
Null	902.39	0	
Linear	894.80	1	
Quadratic	886.28	2	
Local likelihood (span 0.5)	884.65	2.95	
Local scoring (span 0.5)	884.92	2.95	
Restricted cubic spline ($K=3$)	885.57	2	
Restricted cubic spline ($K=4$)	884.65	3	
Restricted cubic spline ($K=5$)	884.46	4	
<i>B</i> -spline with free knots ($d=1, K=1$)	884.32	3	0.005
<i>B</i> -spline with free knots ($d=1, K=2$)	883.47	5	0.023
<i>B</i> -spline with free knots ($d=2, K=1$)	885.98	4	0.032
<i>B</i> -spline with free knots ($d=2, K=2$)	884.03	6	0.056
<i>B</i> -spline with free knots ($d=2, K=3$)	880.17	8	0.041
<i>B</i> -spline with free knots ($d=3, K=1$)	883.82	5	0.027
<i>B</i> -spline with free knots ($d=3, K=2$)	880.70	7	0.029
<i>B</i> -spline with free knots ($d=3, K=3$)	880.00	9	0.063

Table II. Bootstrap results.

Model	95th percentile of <i>p</i> -value distribution	Knot values	Confidence interval
Linear			
One knot <i>B</i> -spline	0.170	47.0	[39.0; 49.7]
Two knots <i>B</i> -spline	0.352	(40.0; 47.0)	[(29.9; 45.2), [42.3; 54.0)]

threshold for heart transplantation. Similar results were obtained by Durrleman and Simon [7] with restricted cubic splines and by Hastie and Tibshirani [6] with local likelihood and local scoring introduction. However, with these methods and with the free knot spline of degree 2 or 3, threshold can be estimated by a minimum or a maximum of the LHR function. Our approach with degree one allows construction of a confidence interval for the knots corresponding to thresholds ([39; 49.7] for one knot).

4.2. Lung cancer data

The fact that there are no large different variations for the log hazard ratio function estimation by a spline with one knot (Figure 3) may indicate that NSE effect is linear, any threshold value can be determinate. Table III indicates that models with spline are not significant ($p > 0.05$), the linear model seems better. An elevated NSE concentration is generally a bad prognostic sign, and this corresponds with the clinical view. This fact can also be confirmed by the large confidence interval [4.0; 39.2] obtained after 1000 replications if we suppose a threshold existence for the NSE. Thus, a patient with a high NSE level has a high risk of death. This suggests that the observation of an increase of NSE at any time during treatment is strongly associated with a worse prognosis.

For the CYFRA, the *p*-value of *B*-spline models are significant (Table IV). Moreover, *p*-value distribution indicates by using the linear spline model with only one knot the existence of one threshold for the CYFRA at 31.5. A confidence interval at 95 per cent for this value is [12.8;

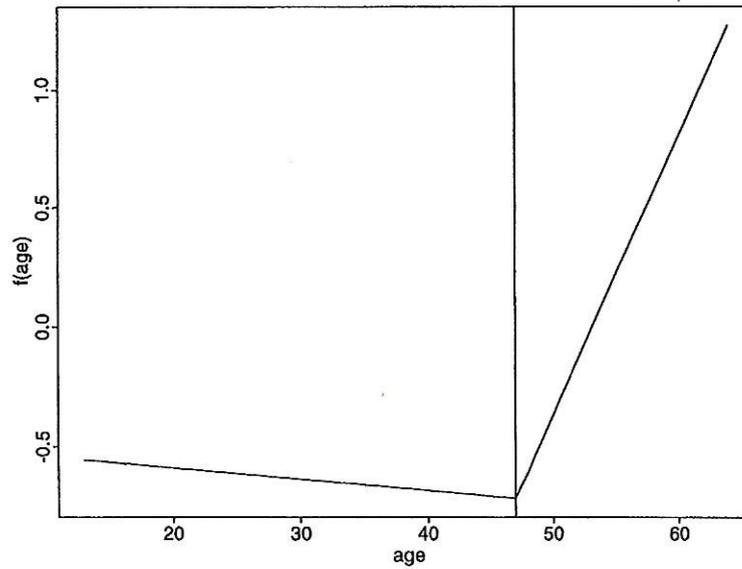


Figure 2. The log hazard ratio function computed with linear spline and one knot for the Stanford heart transplant data.

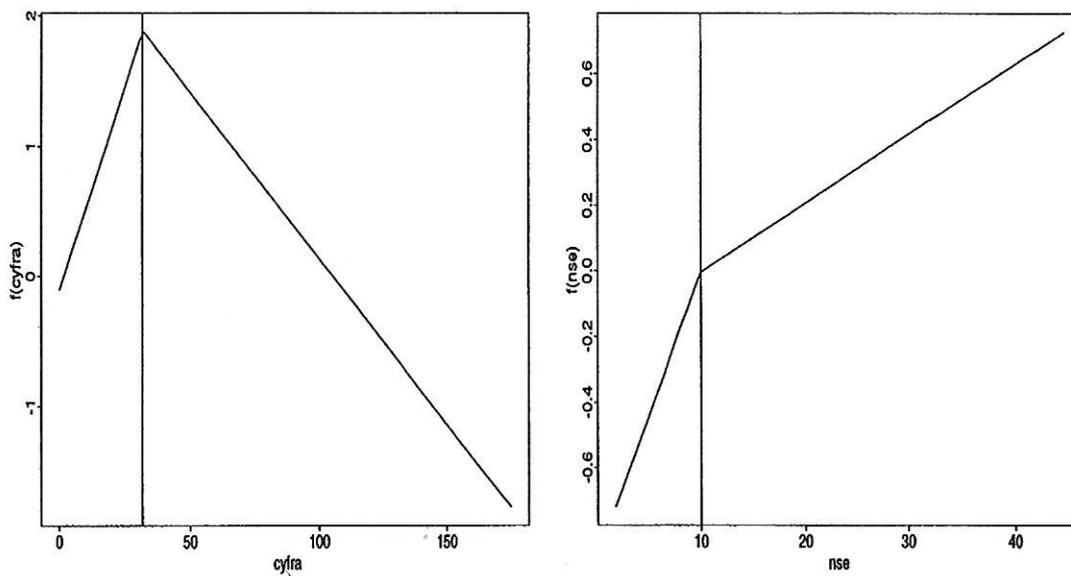


Figure 3. The log hazard ratio function modelled by one knot *B*-spline with CYFRA (left) and NSE (right).

73.1] with a range interval of [0.1;175]. According to Figure 3, threshold presence is clear for CYFRA, because there is an evident break for the log hazard ratio function. Because the hazard function increases then decreases according to the CYFRA values, by using a linear model it does not seem useful to make serial measurements of this marker during treatment (see Boher *et al.*

Table III. Results for the NSE.

Model	-2 log-likelihood	<i>p</i> -value
Linear	804.8	
Free knot spline ($d = 1, K = 1$)	804.0	0.67
Free knots spline ($d = 1, K = 2$)	802.9	0.75
Free knot spline ($d = 2, K = 1$)	803.9	0.83
Free knots spline ($d = 2, K = 2$)	802.5	0.81
Free knot spline ($d = 3, K = 1$)	803.6	0.88
Free knots spline ($d = 3, K = 2$)	802.5	0.89

Table IV. Results for the CYFRA.

Model	-2 log-likelihood	<i>p</i> -value	95th percentile of <i>p</i> -value distribution
Linear	809.5		
Free knot spline ($d = 1, K = 1$)	792.5	2×10^{-4}	0.11
Free knots spline ($d = 1, K = 2$)	790.6	8×10^{-4}	0.15
Free knot spline ($d = 2, K = 1$)	791.3	4×10^{-4}	0.12
Free knots spline ($d = 2, K = 2$)	788.7	9×10^{-4}	0.17
Free knot spline ($d = 3, K = 1$)	790.9	9×10^{-4}	0.16
Free knots spline ($d = 3, K = 2$)	788.4	2×10^{-3}	0.28

[15]). In fact, with a linear estimation, the risk function increases very little with the CYFRA value.

In this example we also present a geometrical approach based on the breaking of the slopes of the linear spline model to test if a knot can effectively be considered as a threshold. Considering only linear spline models, a rather empirical but effective idea to assert if a knot can be interpreted as a threshold, is to examine the difference between the slopes of the piecewise linear LHR function. Denote m_+ (respectively, m_-) the slope of the estimated LHR function on the interval $]\xi, \max(x_i)[$ (respectively, $]\min(x_i), \xi[$). The magnitude of the difference $\mu = m_+ - m_-$ is an indicator of non-linearity in the LHR function. The bootstrapped sample can also be used to estimate the μ -distribution and to test

$$H_0: \xi \text{ is a threshold.}$$

H_0 is rejected if 0 belongs to the 95 per cent- μ -confidence interval. For example, with the lung cancer data with one knot, all μ 's are positive for the CYFRA and only 20 per cent are positive for NSE. These results corroborate that there is no threshold for the NSE. Figure 4 represents LHR functions computed on 50 bootstrapped samples. One can notice the linear effect of the variable NSE in contrast to the variable CYFRA for which a threshold is detected.

5. DISCUSSION

The presented method increases the flexibility of the Cox proportional hazards model for data analysis without having to assume a particular functional form of the considered relationship. It is a mixture of linear and non-linear models in the estimated coefficients.

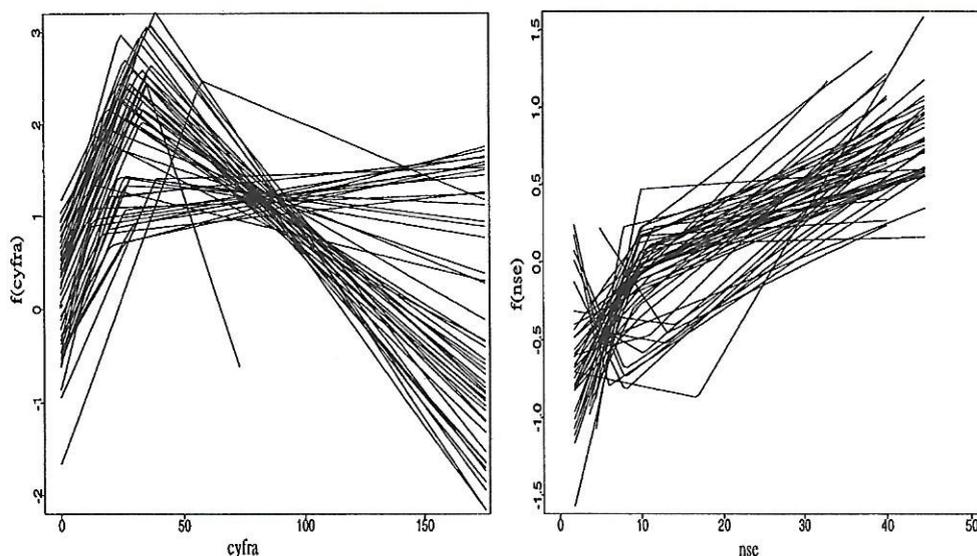


Figure 4. Log hazard ratio functions computed on 50 bootstrapped samples for CYFRA (left) and NSE (right).

When the LHR function changes smoothly, the restricted cubic spline [7] and free knot splines of degree 2 or 3 provide a useful approach for the selection of thresholds that correspond in this case to a minimum or a maximum of the function. Nevertheless, the linear spline model allows the knots to be considered directly as thresholds and provides a confidence interval of these values. Moreover, according to the results presented in Tables I, II and III, splines of degree 3 do not provide better goodness-of-fit values than those from degree 2. To avoid over-fitting effects, we propose using only linear and quadratic splines.

APPENDIX

Proposition 1. The linear model is nested in each spline model.

Proof. We can use another useful basis for splines which is the truncated power basis (see de Boor, reference [8], p. 101) and the same spline of degree d with K distinct knots can be rewritten

$$s(x, \beta, \xi) = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \beta_{d+1} (x - \xi_1)_+^d + \cdots + \beta_{d+K} (x - \xi_K)_+^d$$

where $(x - \xi)_+ := \max\{x - \xi, 0\}$.

With this notation it is clear that the linear model is nested in each spline model of any degree. In fact, suppose only $\beta = (\beta_0, \beta_1, 0, \dots, 0)$ to find the linear function.

ACKNOWLEDGEMENTS

The authors are grateful to Professor J. L. Pujol of Département des Maladies Respiratoires (Hôpital Arnaud de Villeneuve, Montpellier, France) for providing the data on lung cancer and to the referees for helpful comments and suggestions.

REFERENCES

1. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
2. Sleeper LA, Harrington DP. Regressive splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* 1990; **85**:941–949.
3. O'Sullivan F. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal of Scientific and Statistical Computing* 1988; **9**:531–542.
4. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**:78–94.
5. Miller RG, Hallpern J. Regression with censored data. *Biometrika* 1982; **69**:521–531.
6. Hastie T, Tibshirani R. Generalized additive models. *Statistical Science* 1986; **1**:297–318.
7. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989; **8**:551–561.
8. Pujol JL, Grenier J, Daurès JP, Daver A, Pujol H, Michel FB. Serum fragment of cytokeratin subunit 19 measured by CYFRA 21-1 immuno-radiometric assay as a marker of lung cancer. *Cancer Research* 1993; **53**:61–66.
9. de Boor C. *A Practical Guide to Splines*. Springer-Verlag: New-York, 1978.
10. Curry HB, Schoenberg IJ. On Polya frequency functions. IV: The fundamental splines and their limits. *Journal of Analyse Mathematics* 1966; **17**:71–107.
11. MathSoft. *S-Plus version 3.4 for Unix Supplement*. Data Analysis Products Division, MathSoft: Seattle, 1996.
12. Jupp DLB. The Lethargy Theorem, a property of approximation by γ -polynomials. *Journal of Approximation Theory* 1975; **14**:204–217.
13. Owen A. Discussion about multivariate adaptative regression splines. *Annals of Statistics* 1991; **19**:102–112.
14. Feder PI. On the likelihood ratio statistic with applications to broken line regression. PhD dissertation, Department of Statistics, Sanford University, 1967.
15. Boher JM, Pujol JL, Grenier J, Daurès JP. Markov model and markers of small cell lung cancer: assessing the influence of reversible serum NSE CYFRA 21-1 and TPS levels on prognosis. *British Journal of Cancer* 1999; **79**:1419–1427.