Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions

Harald Heinzl *, Alexandra Kaider

Department of Medical Computer Sciences, University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

Received 27 September 1996; received in revised form 29 May 1997; accepted 25 July 1997

Abstract

The Cox proportional hazards model is the most popular model for the analysis of survival data. The use of cubic spline functions allows investigation of non-linear effects of continuous covariates and flexible assessment of time-by-covariate interactions. Two main advantages are provided—no particular functional form has to be specified and standard computer software packages like SAS or BMDP can be used. A SAS macro which implements the method is presented. © 1997 Elsevier Science Ireland Ltd.

Keywords: Proportional hazards regression model; Non-linear functional relationship; Time-by-covariate interaction; Binary time-dependent covariate; Restricted cubic spline function; SAS macro

1. Introduction

In the early stages of scientific investigation the exploratory aspect of a data analysis comes in the foreground. Nonparametric methods like cubic regression splines can be used for exploring data. Cubic regression splines are flexible tools for identifying hidden relationships, produce visibly smooth curves, are easy to interpret, allow the testing of statistical hypotheses, and can be computed using widely available software packages like SAS (SAS Institute, Cary, NC) and BMDP (BMDP Statistical Software, Los Angeles, CA). It is straightforward to apply them [1–3] to improve the flexibility of the Cox proportional hazards regression model [4,5], which is the standard tool for assessing prognostic factors in clinical studies with a failure time outcome variable.

Besides all their favorable features the practical use of cubic regression splines is limited by their rather bulky formulas. The following SAS macro RCS is a mean to overcome this problem. Controlled by some simple statements the macro uses PUT statements to generate SAS code (i) to apply cubic splines in the Cox model (PROC PHREG), (ii) to prepare the results for plotting (PROC IML), and (iii) to plot them (PROC GPlot). That is, the user gets an executable SAS program, whose details can be easily modified if required (for instance, change details of the graph, add a subtitle, etc.).

Section 2 contains a brief definition of restricted cubic splines. The problems we want to solve with them are described in Section 3. The RCS-macro is introduced in Section 4.

2. Definition of restricted cubic splines

The expression for a restricted (or natural) cubic spline function with \( k \) knots, \( t_1 < \ldots < t_k \), is given by

\[
C(u) = \beta_0 + \sum_{j=1}^{k-2} \theta_j C_j(u),
\]

where \( C_1(u) \ldots C_{k-2}(u) \) are cubic terms,

\[
C_j(u) = \frac{(u-t_j)^3_+}{[t_k-t_{j-1}]} + \frac{(u-t_k)^3_+}{[t_k-t_{j-1}]}, \quad j = 1 \ldots k-2.
\]

The notation \(( \cdot )_+\) is very common in the splines literature, and stands for \((a)_+ = \max(0, a)\). A more detailed description of the development of restricted cubic splines may be found in [1–3].

Note that,

- \( C(u) \) has continuous first and second derivatives;
- \( C(u) \) is linear in \( u \) for \( u < t_i \) and \( u > t_k \) ("linear in the tails");
- although \( C(u) \) looks rather complicated, it is a linear function with regard to the \( k \) parameters \( \beta_0, \beta_1, \ldots, \beta_{k-2} \) and readily available statistical packages for linear models can be used to estimate them (for instance, PROC PHREG of SAS or procedure 2L of BMDP for Cox regression);
- \( C(u) \) is linear in the parameters which enables the use of standard tools of statistical inference.

Often splines are used to approximate an unknown, smooth function \( f(u) \). Now, a test that \( \beta_1 = \ldots = \beta_{k-2} = 0 \) is equivalent to a test that \( f(u) \) is constant, and a test that \( \theta_1 = \ldots = \theta_{k-2} = 0 \) is equivalent to a test that \( f(u) \) is linear, see [1–3].

The graphical representation of the estimated function \( \tilde{C}(u) \) is the main part of the output, as the estimated parameters alone are usually hard to interpret. To investigate the plot of \( \tilde{C}(u) \) against \( u \) it is often helpful to add confidence bands for \( \tilde{C}(u) \). Due to the simple linear structure of our model these can be easily computed. For a fixed value \( u_0 \), \( \tilde{C}(u_0) \) can be written as

\[
\tilde{C}(u_0) = \tilde{\beta}' U_0,
\]

where \( \tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_{k-2})' \) and \( U_0 = (1, u_0, C_1(u_0), \ldots, C_{k-2}(u_0))' \). If \( V \) is the large-sample covariance matrix for \( \tilde{\beta} \), then a \( 1 - \alpha \) confidence interval for \( \tilde{C}(u_0) \) is given by

\[
[\tilde{C}_{low}(u_0), \tilde{C}_{upp}(u_0)] = \tilde{\beta}' U_0 \mp (\gamma V U_0 V')^{1/2},
\]

where \( \gamma = \chi^2_{p,1-\alpha} \) is the \((1-\alpha)\)-quantile of the \( \chi^2 \) distribution on \( p \) degrees of freedom. When \( p \) is set to the number of covariates this method yields Scheffé-type simultaneous confidence bands, and when \( p \) is set to 1 it yields standard pointwise confidence bands [2].

The number and the position of the knots of a cubic spline function must be specified. With sample sizes common for clinical trials empirical evidence [1–3] suggests that 3–5 knots will usually suffice. The knots should be placed 1. at quantiles of the observed distribution of \( u \); 2. near but not at the extremes; 3. roughly uniform over the quantiles.

Reasonable sets of percentiles for knot location seem to be \{5, 50, 95\}, \{5, 25, 75, 95\} and \{5, 25, 50, 75, 95\} for \( k = 3 \), 4 and 5 knots, respectively. The use of prespecified quantiles makes knot selection less subjective, and data analyses based on commonly accepted automatic knot placement schemes can be regarded as more ob-
jective and comparable. For more details on knot location see [1-3]. Note that the result of fitting a cubic spline to a data set will usually be insensitive to the location of knots unless they occur in an extremely non-uniform way over the covariate space [1,3].

3. The problem

We assume that data from a clinical trial of $n$ patients, $(y_1, s_1, z_1)\ldots(y_n, s_n, z_n)$, are available. Thereby $y_i$ denotes the observed survival time, the interval for which the $i$th patient has been observed from entering the study till leaving. The binary status variable $s_i$ indicates, whether leaving has been through failure like death, relapse or infection ($s_i = 1$) or through censoring ($s_i = 0$). The covariable $Z$, which we are interested in, is unspecified for the present.

3.1. Assessing the non-linear effect of a continuous covariable $Z$

Let $z_1\ldots z_n$ be realisations of a continuous prognostic factor $Z$ like age (years), serum bilirubin (mg/dl), or systolic blood pressure (mmHg). When fitting a Cox model we assume the hazard of failure at time $t$ for the $i$th patient to be

$$h(t; z_i) = h_0(t) \exp(\beta z_i)$$

The unknown parameter $\beta$ is estimated by partial likelihood [4,5], i.e. no further assumptions about the unknown baseline hazard function $h_0(t)$ are imposed. It follows for this model that the log hazard ratio function (LHR) with respect to $Z$ is a linear function of $Z$:

$$\text{LHR}(Z) = \log(h(t; Z)/h_0(t)) = \beta Z$$

Although $h_0(t)$ and $h(t; Z)$ are unknown, we suppose that a unit change in $Z$ has the same effect on the patients LHR all over the range of $Z$. This type of modelling is restrictive, since the behavior of $\text{LHR}(Z)$ may be highly non-linear. To explore the nature of the $\text{LHR}(Z)$ we need more flexibility to model a continuous covariate. Therefore, let

$$h(t; Z) = h_0(t) \exp(f(Z))$$

which yields

$$\text{LHR}(Z) = \log(h(t; Z)/h_0(t)) = f(Z)$$

Since the baseline hazard $h_0(t)$ corresponds to $h(t; Z = 0)$, $f(Z)$ must be equal to zero for $Z = 0$. We have to be aware of this fact when using cubic splines to approximate $f(Z)$:

$$\text{LHR}(Z) = \log(h(t; Z)/h_0(t)) \approx C(Z) - C(0) = \beta Z + \sum_{j=1}^{k-2} \theta_j (C_j(Z) - C_j(0)).$$

Note, if the smallest knot is non-negative, $t_1 \geq 0$, then $C_j(0)$ will equal zero, $j = 1\ldots k - 2$.

By way of example, let $Z$ be patient-age in years. It may be not be interesting nor sensible to report hazard ratios relative to the age of zero years, but it may be of great interest to report them in relation to the age of, say 35 years. Therefore, the LHR function for a given reference value $m$ is

$$\text{LHR}_m(Z) = \log(h(t; Z)/h(t; m)) = f(Z) - f(m) \approx C(Z) - C(m) = \beta (Z - m) + \sum_{j=1}^{k-2} \theta_j (C_j(Z) - C_j(m)).$$

Note that once we have fitted $\beta$, $\theta_1\ldots \theta_{k-2}$, we can easily change the reference value $m$, since there is only a shift in the axis of ordinates. Confidence bands for the estimated LHR function with reference value $m$ can be constructed by applying the same principle as for $C(u)$ in Section 2.

3.2. Assessing interactions of time with a covariable $Z$

Let $z_1\ldots z_n$ be realisations of a prognostic factor $Z$, which is either dichotomous or continuous. It is in the definition of a proportional hazards model that the LHR function with respect to time $t$ is constant:

$$\text{LHR}(t) = \log(h(t; Z)/h_0(t)) = \beta Z$$
In contrast to Section 3.1 we are now considering the LHR function as function of time $t$. The functional form of $\text{LHR}(t)$ can be explored in a flexible way by using the fact that time-dependent covariates can be included in the Cox model in a straightforward manner. Therefore, let

$$\text{LHR}(t) = \log(h(t;Z)/h_0(t)) = g(t)Z$$

and approximate $g(t)$ with cubic splines. When considering a unit change in $Z$,

$$\text{LHR}_+(t) = \log(h(t;Z = z_0 + 1)/h(t;Z = z_0)) = g(t) \approx C(t) = \beta_0 + \beta_1 t + \sum_{j=1}^{k-2} \theta_j C_j(t)$$

Note that the knots have to be non-negative, $0 \leq t_1 < \ldots < t_k$.

The estimated log hazard ratio over time for a unit change in $Z$ is simply $\hat{C}(t)$.

3.3. Assessing interactions of time with a binary time-dependent covariable $Z$

Let $z_1, \ldots, z_n$ contain the times of occurrence of a specific binary time-dependent event of interest, i.e. at time zero when entering the study the $i$th patient starts in group 0 and changes to group 1 at time $z_i$. If this particular time-dependent event does not occur during follow-up, $z_i$ contains a missing value and the patient remains in group 0 until leaving. A famous example for a binary time-dependent covariate is provided by the Stanford Heart Transplant Programme. Patients entered the study, when they had been accepted for heart transplantation. Whenever a donor heart became available, medical judgement was used to select the patient who should receive it. That is, a patient started the study as non-transplanted (group 0) and after some waiting time $z_i$ he became transplanted (changed to group 1). Some patients died before a donor heart was available for them.

A very natural and straightforward clinical question is to ask about the effect of this time-dependent event on survival. Usually such a question is answered by defining a binary time-dependent covariate $x_i(t), i = 1 \ldots n$:

$$x_i(t) = \begin{cases} 0 & 0 < t < z_i \\ 1 & t \geq z_i \end{cases}$$

and assuming a Cox proportional hazard model

$$h_i(t) = h_0(t) \exp(\beta x_i(t))$$

The baseline hazard $h_0(t)$ is the hazard function for a patient, who stays in group 0 all the time. The LHR function for the $i$th patient is constant over time for $t \geq z_i$ (notice that by definition $\text{LHR}_i(t)$ is equal to zero and therefore constant over time for $t < z_i$, also), $i = 1 \ldots n$:

$$\text{LHR}_i(t) = \log(h_i(t)/h_0(t))$$

This type of modelling 'by jump' is restrictive, since the behavior of $\text{LHR}_i(t)$ after the occurrence of the time-dependent event may be (i) unchanged for the present (the 'jump' is long in coming), (ii) smoothly increasing or decreasing, (iii) of a temporary nature only (transient), (iv) alternating, or (v) a combination of these or other patterns. To become more flexible the hazard for the $i$th patient can be modified by

$$h_i(t) = h_0(t) \exp(l(t - z_i)),$$

where $l(u) = 0$ for $u < 0$. The LHR function for the $i$th patient becomes

$$\text{LHR}_i(t) = \log(h_i(t)/h_0(t)) = l(t - z_i)$$

Restricted cubic splines can be used to approximate the unknown functional form of $l(u)$ for $u \geq 0$. That is,

$$\text{LHR}_i(t) = l(t - z_i) \approx x_i(t)C(t - z_i)$$

$$= \beta_0 x_i(t) + \beta_1 (t - z_i) + \sum_{j=1}^{k-2} \theta_j C_j(t - z_i).$$

Note that here the knots have to be non-negative, also, $0 \leq t_1 < \ldots < t_k$.

By definition each patient has its own LHR function. When plotting the estimated LHRs it suffices to plot their common part after the occurrence of the binary time-dependent event, which is $\hat{C}(u)$, for $u \geq 0$. 


4. Program description

A SAS macro has been written which generates a SAS program to implement cubic splines in the Cox model. It requires the existence of a SAS input data set containing a survival time variable, a status variable, and up to 20 covariates. The user has to pass the names of these variables to the macro. To avoid confusion with auxiliary variables of the procedure PHREG, no variable name in the input data set should start with a double underscore.

The user can specify, which of the covariates should enter the analysis as ordinary covariates and which of them should be modelled with cubic splines, respectively. If the latter is chosen for a certain covariate, the user has to decide whether a non-linear functional relationship, a time-by-covariate interaction, or a time by binary time-dependent covariate interaction should be investigated. Additionally, the knot positions have to be provided.

The RCS-macro generates a SAS program, which consists of a PHREG procedure for estimating the Cox model, and as many IML and GPLOT procedures as there are spline modelled covariates. The IML procedures are used to prepare the estimated results before plotting them.

```sas
%RCS(
    TITLE=%STR(acute leukemia dataset as printed in Hess, 1994),
    DATA=gehan,   DIRDATA=%STR(b:\hes\data),
    PROGRAM=%STR(E:\study\results\kenneth.sas),
    TIME=svmtime, STATUS=svmtime,
    CV1=GROUP,   WHAT1=1,   KNOTS1=6 10 19,
    TIMEUNIT=weeks
);
```

with the GPLOT procedures. The SAS program has to be submitted by the user to get the actual results.

The chosen strategy yields a 2-fold advantage for the user. On one hand, a potential source of error due to writing down bulky formulas is removed, since these formulas will be produced automatically. On the other hand, the user keeps control of the SAS program, and changes can be made if desired, especially details of the graphical output are predestined for alteration.

The SAS macro RCS, a detailed description for its use, and three sample data sets are available via world wide web at ‘http://www.akh-wien.ac.at/imc/biometrie/rcs.zip’.

4.1. Example

Hess [2] presented a well-known 2-sample dataset for 42 patients with acute leukaemia. A new treatment, 6-mercaptopurine (6-MP), was compared to placebo in the maintenance of remissions. SURVTIME (remission time in weeks), SURVSTAT (1 = end of remission; 0 = censored) and GROUP (0 = placebo; 1 = 6-MP) are the only variables in the data set. The questions of interest are for the effect of the 6-MP drug, and if this effect depends on time. The latter question can be reformulated in: is the proportional hazards assumption valid or is there an interaction of time-by-covariable GROUP? As outlined in Section 3.2 cubic spline functions provide an exploratory tool to tackle this problem in a reasonable way. To generate a SAS program for fitting a 3-knots cubic spline function for testing a possible time by GROUP interaction, the following statements of the RCS-macro can be used. Note that for specifying more complicated character strings the use of the %STR macro function is recommended.

The meaning of the various statements is as follows.

- **TITLE**: title of the analysis
- **DATA, DIRDATA**: SAS data set and input directory which contains the input data
- **PROGRAM**: external filename for the SAS program file to generate; both directory and filename have to be specified
- **TIME, STATUS**: survival time and status variable; censored observations must be coded as 0, failures can be coded with arbitrary values as long as they are non-zero
COV1: name of the covariate; in case that there are more covariates, we would use COV2, COV3 and so on
WHAT1: what should happen with COV1; the chosen option requests modelling of a time-by-
covariate interaction with cubic splines
KNOTS1: specifies strictly increasing knot-positions for spline-modelling of COV1; use blanks
as separators
TIMEUNIT: label of X-axis for graphical output

Submitting the statements above will generate the output file \"e:\study\results\kenneth.sas\". Here is
its contents (PROC IML code has been shortened):

LIBNAME _DATA 'B:\HESS\DATA\';

TITLE 'ACUTE LEUKAEMIA DATASET AS PRINTED IN HESS, 1994';

PROC PHREG DATA=_DATA.GEHAN COVOUT OUTTEST=_RCS;
MODEL SURVTIME*SURVSTAT(0) = GROUP __1_LIN__1_1 /RL;

************ spline modelling of the time-dependent effect;
************ of fixed covariate GROUP;
************ with 3 knots located at:
************ 6 10 19;
__1_1=((SURVTIME-6)**3)*(SURVTIME>6)
-((SURVTIME-10)**3)*(SURVTIME>10)
*(10-6)/(19-10)
+((SURVTIME-19)**3)*(SURVTIME>19)
*(10-6)/(19-10);
__1_LIN=GROUP*SURVTIME;
__1_1=GROUP*__1_1;

*----------- Testing variable: GROUP -----------;
EFFECT1: TEST GROUP, __1_LIN, __1_1;
NONCON1: TEST __1_LIN, __1_1;
NONLIN1: TEST __1_1;
RUN;

*-------------------- End of PROC PHREG ---------------------;

*-------------------- Graph for GROUP --------------------;
PROC IML;
NPOINTS=101; * Number of points to build the graphic;
LOWEREND=0; *Smallest value for X-axis;
UPPEREND=19; *Largest value for X-axis;

(... code omitted ...)
CREATE _RCS1 VAR { F FE Z X }; APPEND; CLOSE _RCS1;
QUIT;

SYMBOL1 C=RED L=1 I=JOIN WIDTH=5;
SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

PROC GPLOT DATA=_RCS1;
PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
TITLE2 'GROUP';
LABEL X=WEKKS;
LABEL F=LOG HAZARD RATIO;
RUN;
Fig. 1. LHR function (and 95% pointwise confidence band) for acute leukaemia data estimated by a cubic spline function with 3 knots at 6, 10, and 19 weeks.

The chosen cubic spline has 3 knots placed at 6, 10, and 19 weeks. The PHREG code for estimating the Cox model has three regressors, GROUP, _1_LIN and _1_1. The latter is a spline function in time defined in the programming section of the PHREG procedure.

_1_LIN represents a linear time term and is defined in the programming section, also. Both _1_LIN and _1_1 are multiplied with GROUP, since we are modelling the time-dependent effect of a unit change in the covariate of interest.

The TEST-statement labelled EFFECT1 produces a Wald test for the null hypothesis that the regression coefficients of all variables related to GROUP are equal to zero, which is equivalent to the null hypothesis that GROUP has no effect at all. The NONCON1 (NONLIN1) test is a test for the regression coefficients of the spline variables including (without) the linear term, which is equivalent to the null hypothesis that the effect of GROUP on survival is constant (linear), respectively.

The PROC IML program prepares the data for the PROC GLOT. Note, that there are three default settings at the beginning, which can be changed by the user, if he wishes to do so. NPOINTs determines the number of points, of which the plotted lines will be made of; default value is 101. LOWEREND and UPPEREND determine the range of the plotted X-axis; 0 and the largest knot position will be used as default values, respectively. Finally, the GPLOT procedure is used to plot the estimated spline function and its pointwise 95% confidence band versus time. This corresponds to an estimate of the LHR function of a unit change in GROUP. Since GROUP is dichotomous (0 = placebo; 1 = 6-MP), a unit change represents the difference between the two therapies.

Submitting the SAS-code of 'c:s\study\ results\kenneth.sas' will produce a SAS-listing with PROC PHREG output and a graph of the estimated cubic spline function (Fig. 1). Parts of the SAS-listing (results of Wald tests) are shown below:
ACUTE LEUKAEMIA DATASET AS PRINTED IN HESS, 1994

The PHREG Procedure

Data Set: __DATA.GEHAN
Dependent Variable: SURVTIME
Censoring Variable: SURVSTAT
Censoring Value(s): 0
Ties Handling: BRESLOW

(... output omitted ...)

Linear Hypotheses Testing

<table>
<thead>
<tr>
<th>Label</th>
<th>Wald</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFFECT1</td>
<td>13.4877</td>
<td>3</td>
<td>0.0037</td>
<td></td>
</tr>
<tr>
<td>NONCON1</td>
<td>0.4239</td>
<td>2</td>
<td>0.8090</td>
<td></td>
</tr>
<tr>
<td>NONLIN1</td>
<td>0.4134</td>
<td>1</td>
<td>0.5203</td>
<td></td>
</tr>
</tbody>
</table>

From the EFFECT1-test we can conclude that there is a significant difference in the duration of remission between therapy 6-MP and placebo. From the NONCON1-test there is no evidence that this effect is time-dependent, although the power of this test may be limited by the sample size of 42 patients. The graphical representation of the results (Fig. 1) confirms the NONCON1-test result of a fairly constant hazard ratio over time, but be aware that the estimated function is a straight line until the first knot at week 6 by definition.

At the end of Section 2 we outlined a strategy for knot placement recommended by several authors. According to it, Hess [2] suggests to use 2, 10, and 32 weeks for the data set in question. It is simple to adapt the call of the RCS-macro described before to this new choice. We just have to change the KNOTS1-statement, although for purpose of clearness we usually would change the external filename for the SAS program file and the title of the analysis, also.

References