Fitting latency models using B-splines in EPICURE for DOS

Michael Hauptmann, Jay Lubin

January 11, 2007

1 Introduction

Disease latency refers to the interval between an increment of exposure and a subsequent change in an individual's risk. This implies that the risk from a certain exposure history does not depend on cumulative exposure alone, but on the timing of exposure. It can be expected that risk varies over time in a smooth way, and this variation can be described by a latency curve.

We explain how to apply a spline function model to data on exposure history and disease. Splines are piecewise polynomial functions [1]. Theoretically, latency patterns could be described by estimating separate risk parameters for exposures received in each year prior to current age. However, the number of parameters would be large, and all the parameters could not be estimated due to limited data and possibly correlations between exposures in subsequent time intervals. The large number of parameters that would have to be estimated in such a nonparametric approach can be reduced by the use of spline functions. The use of cubic splines is a mild restriction because of their flexibility in approximating smooth functions. The model that includes risk with total cumulative exposure, which corresponds to a constant latency curve, is nested in the spline model. This method has been applied in occupational lung cancer epidemiologic studies where the exposures were asbestos [2] and radon [3]. Other approaches to analyze latency in epidemiologic studies include simple exploratory techniques [4, 5] or the bilinear model developed by Langholz et al. [6].

We briefly describe the spline function model and explain the use of EPICURE code to fit this model.

2 The spline latency model

Let $x(t)$ be the exposure during the year from $t-1$ to t years prior to the death of a case or the corresponding age for a control. For example, for an individual with attained age 52.4, $x(1)$ is the exposure from age 51.4 through 52.4, and so forth.

Thus, $x(1), \ldots, x(40)$ represent the exposure history, and $\sum_{t=1}^{40} x(t)$ is total cumulative exposure. We start with a general model for the relative risk (RR), $\mathrm{RR} = 1 + \sum_{t=1}^{40} \theta_t x(t)$, where $\theta_1, \ldots, \theta_{40}$ are parameters that fully describe the latency curve. In general, data will be insufficient to estimate the full set of parameters, $\theta_1, \ldots, \theta_{40}$. Our approach is then to apply mild constraints to the θ_t's and estimate a functional form that describes their behavior. Suppose $\mathrm{RR} = 1+\sum_{t=1}^{40} s(t;\theta)x(t)$, where $s(t;\theta)$ is a function of time t and a parameter vector θ that models the year-specific ERR per unit exposure, i. e., $s(t;\theta)$ is the ERR per unit exposure received t years in the past. The weighted sum $\sum_{t=1}^{40} s(t;\theta)x(t)$ represents the ERR for the exposure profile $x(1), \ldots, x(40)$ compared to a zero profile, i. e., a non-exposed individual.

A cubic B-spline is used to model $s(t;\theta)$ [7]. Splines are smooth (i. e., continuously differentiable) piecewise polynomial functions. They are segmented by interior knots. Cubic splines have certain optimum properties for the approximation of curves [1]. The parameters cannot be interpreted directly, but the estimated spline function and corresponding confidence intervals can be plotted.

Spline models with different number and placement of knots are not nested. Therefore, the number and placement of knots cannot be evaluated by likelihood ratio tests. To assure a smooth curve and to avoid overfitting, cubic splines with a small number of interior knots (three or less) may be considered. Two approaches can be applied to determine the placement of knots. A profile likelihood search can be performed for one interior knot by evaluating the deviance of models for a series of possible knot locations. This is computationally cumbersome for multiple knots. Alternatively, knot positions can be selected such that the study population accumulated approximately constant proportions of its cumulative exposure between two adjacent knots. For a cubic spline with one interior knot, e. g., 5 spline parameters have to be estimated and the knot position has to be determined.

The simple linear excess relative risk (ERR) model in cumulative exposure is included in the spline model when the function is constant over time, that is $s(t;\theta) = \beta$ for all t. In this case, β is the ERR per unit exposure. A likelihood ratio test can be performed to test if the data are consistent with no variation in the year-specific risk, i. e., cumulative exposure. The details are as follows.

The function $s(t;\theta)$ is modeled using a B-spline as described by de Boor [1]. A spline of order k on the interval $[a, b]$ consists of polynomials of order k on the $m + 1$ segments defined by m inner knots $a < t_1 < \ldots < t_m < b$. Adjacent polynomials are smoothly joined, so that first and second derivatives agree at the knots.

Using a numerically favorable representation of splines, the space of splines can be spanned with $m+k$ basis functions $B_i(t)$, called B-splines. The knot list has to be augmented by six associated arbitrary "slack" knots. Without loss of generality, let $t_{-(k-1)} = b - (k - 1)$, $t_{-(k-2)} = b - (k - 2)$, ..., $t_0 = b$ and $t_{m+1} = b$, $t_{m+2} = b + 1$, ..., $t_{m+k} = b + k - 1$. Using a differently augmented knot list results in different basis vectors that nonetheless span the same space and have the same properties.

Starting with $B_{i,1}(t) = 1$ if $t_i \leq t < t_{i+1}$ and zero otherwise, the B-spline basis functions are defined by the recurrence relation

$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t).$$

The spline function has the form $s(t;\theta) = \sum_{i=-(k-1)}^{m} \theta_i B_{i,k}(t)$. The spline parameters can be estimated by maximizing the likelihood function using EPICURE [8] code.

To estimate the position of one interior knot by a profile likelihood search, the likelihood function has to be evaluated for the series $a+1, a+2, \ldots, b-1$ of possible locations of the single interior knot. It has to be noted that the pointwise confidence intervals for the estimated spline function do not include the added variability from estimating the knot position.

Alternatively, for m inner knots and thus $m+1$ intervals knot locations can be chosen such that each interval includes $1/(m+1) \times 100$ percent of the total cumulative study population exposure. More precisely: the jth knot t_j is chosen so that $t_j = \max\{t = a, \ldots, b \,|\, \sum_{i=1}^{n} \sum_{\ell=1}^{t} x_i(\ell) / \sum_{i=1}^{n} \sum_{\ell=1}^{T} x_i(\ell) \leq (j-1)/(m+1)\}$, where n is the number of subjects in the study.

For deriving variance formulae for the risk parameter as well as for the weights we switch to matrix notation. Omitting the order k from the index of the B-spline basis functions, $B_{i,k}(t) = B_i(t)$, a B-spline is then represented by the collocation matrix

$$B = [B_1, \ldots, B_m] = \begin{bmatrix} B_1(a) & \ldots & B_m(a) \\ \vdots & & \vdots \\ B_1(b) & \ldots & B_m(b) \end{bmatrix}_{b-a+1 \times m.}$$

The spline function is given by

$$s(t;\theta) = [B_1(t), \ldots, B_m(t)]\theta = B(t)_{1\times m}\theta_{m\times 1}$$

and the values of the spline function at all time points between a and b are contained in

$$s = [s(a;\theta), \ldots, s(b;\theta)]' = B\theta.$$

The variance of the spline function values is given by

$$\mathrm{Var}(s) = \mathrm{Var}[B\theta] = B\mathrm{Var}[\hat{\theta}]B',$$

where $\mathrm{Var}[\hat{\theta}]$ is the variance-covariance matrix of the vector of maximum likelihood estimates $\hat{\theta}$.

3 Program structure

The program consists of the following files

- calculation of spline covariates
  `bspline.cmd`: main file for calling others
  `arrays.cmd`: sets arrays, called by `bspline.cmd`
  `knots.cmd`: augments the given knot list, called by `bspline.cmd`
  `bx1.cmd`: part 1 of recursion algorithm, called by `bspline.cmd`
  `bx2.cmd`: part 2 of recursion algorithm, called by `bx1.cmd`

- calculation of Wald confidence intervals
  `ci_wald.cmd`: main file for calling confidence interval calculation files
  `ci_wald1.cmd`: part 1 of recursion algorithm, called by `ci_wald.cmd`
  `ci_wald2.cmd`: part 2 of recursion algorithm, called by `ci_wald1.cmd`

- calculation of likelihood ratio confidence intervals
  `ci_lr.cmd`: main file for calling confidence interval calculation files
  `ci_lr1.cmd`: part 1 of recursion algorithm, called by `ci_lr.cmd`
  `ci_lr2.cmd`: part 2 of recursion algorithm, called by `ci_lr1.cmd`

- calculation of bootstrap confidence intervals
  `ci_btsp.cmd`: main file for calling confidence interval calculation files
  `ci_btsp1.cmd`: part 1 of recursion algorithm, called by `ci_btsp.cmd` and `ci_btsp3.cmd`
  `ci_btsp2.cmd`: part 2 of recursion algorithm, called by `ci_btsp1.cmd`
  `ci_btsp3.cmd`: draw bootstrap samples, fit weight function model and calculate weight function, called by `ci_btsp.cmd`

`bspline.cmd` has to be edited to provide the program with the knots and the order of the spline to fit, and the name of the exposure profile variables. Running `bspline.cmd` performs some introductory operations like intializing arrays (`arrays.cmd`) and augmenting the knot list (`knots.cmd`). Eventually, the spline covariates `bx1`, `bx2`, ... are calculated. Only the first m of these are needed. Those not needed are set to zero. After using these variables in a model as described below, Wald confidence intervals are calculated by running script `ci_wald.cmd`. This script calls `ci_wald1.cmd` and `ci_wald2.cmd` and writes estimated weights with standard errors and confidence intervals for each time point `a`, `a+1`, ..., `b` to the report file `ci_wald.txt`. These yearly log odds ratios (and confidence limits) can be plotted against time. Similar code is available to calculate likelihood ratio and bootstrap confidence intervals.

4 How to use the program?

First, copy the files to a directory and change the paths in the `WHILE` statements to that directory.

Start any EPICURE module and load a data set that contains an exposure profile on time since exposure scale, i. e. variables `x1`, `x2`, ..., `x50`, where `x10`, say, is exposure in year 10 before interview. Edit script `bspline.cmd` and provide the end years `#a` and `#b` of the time interval you are interested in and the interior knots `#knot1`, `#knot2`, ... Also provide the number of interior knots and

the order of the spline you want to fit (2, 3, 4 for a linear, quadratic, or cubic spline, respectively). Then identify the exposure history variables by editing the `ARRAY` statement. Do all that in the head of the `bspline.cmd` script in the appropriate places.

Run `bspline.cmd`. It creates spline covariates `bx1`, ..., `bx12`. Only the first m are nonzero. Fit a risk model that includes all variables `bx1`, ..., `bx12` in a way that the associated parameters have numbers 2, 3, ..., 13. This should be possible for almost all models. An intercept `%CON` is usually automatically added as parameter no. 1. If the model is stratified or a conditional regression model, the intercept is not needed and should be fixed at zero using the command `PARA 1=0@` in order to avoid that one of the spline covariates `bx1`, ..., `bx12` is being aliased. You may need to specify the additive risk model using `RRISK ADD@`. After fitting the model, run script `ci_wald.txt`. This script will automatically pick the parameters of the spline covariates (numbers 2, ..., 13) and calculate estimates, standard errors and approximate normal confidence intervals for the yearly log odds ratios, which are a linear combination of the parameters and the B-splines, by calculating the B-splines again and using the `LINCOMB` command. Note that for technical reasons the current model has to include parameters number 1 through 13 even if there are less spline variables. Nuisance parameters can be set to zero.

The results for each time point t between `#a` and `#b` are written to the report file `ci_wald.txt` and can be plotted, e. g. using MATLAB programs `splot.m` and `stdplot.cmd` (the latter is for plotting confidence intervals).

The file `aae2tse.cmd` contains EPICURE code to transform annual exposures on the age-at-exposure scale to the latency scale.

Suggestions are always welcome. Send them to `m.hauptmann@nki.nl`.

References

[1] C. de Boor. *A practical guide to splines*. Number 27 in Applied Mathematical Science. Springer, New York, 1978.

[2] M. Hauptmann, K. Berhane, B. Langholz, and J. H. Lubin. Using splines to analyse latency in the Colorado Plateau uranium miners cohort. *J Epidemiol Biostat*, 6:417–424, 2001.

[3] M. Hauptmann, H. Pohlabeln, J. H. Lubin, K. H. Jöckel, W. Ahrens, I. Brüske-Hohlfeld, and H. E. Wichmann. The exposure-time-response-relationship between occupational asbestos exposure and lung cancer in two German case-control studies. *Am J Ind Med*, 41:89–97, 2002.

[4] M. M. Finkelstein. Use of time windows to investigate lung cancer latency intervals at an Ontario steel plant. *Am J Ind Med*, 19:229–235, 1991.

[5] M. auptmann, J. H. Lubin, P. S. Rosenberg, J. Wellmann, and L. Kreienbrock. The use of sliding time windows for the exploratory analysis of tem-

poral effects of smoking histories on lung cancer. *Stat Med*, 19:2185–2194, 2000.

[6] B. Langholz, D. C. Thomas, A. Xiang, and D. Stram. Latency analysis in epidemiologic studies of occupational exposures: applications to the Colorado Plateau uranium miners cohort. *Am J Ind Med*, 35:246–256, 1999.

[7] M. Hauptmann, J. Wellmann, J. H. Lubin, P. S. Rosenberg, and L. Kreienbrock. The analysis of exposure-time-response relationships using a spline weight function. *Biometrics*, 56:1105–1108, 2000.

[8] D. L. Preston, J. H. Lubin, D. A. Pierce, and M. E. McConney. *Epicure Release 2.0*. HiroSoft International Corporation, Seattle, 1996.