PHMPL: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data.

Pierre Joly, Luc Letenneur, Ahmadou Alioum, Daniel Commenges

PHMPL: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data

Pierre JOLY, Luc LETENNEUR,

Ahmadou ALIOUM, Daniel COMMENGES

ISPED

Université Victor Segalen Bordeaux II

146, rue Léo Saignat

33076 Bordeaux Cedex, France

Tel : (33) 5 57 57 11 36

Fax : (33) 5 56 99 13 60

E-mail: Pierre.joly@dim.u-bordeaux2.fr

Corresponding author : Pierre JOLY

October 25, 2007

Abstract

The Cox model is the model of choice when analyzing right-censored and possibly left-truncated survival data. The present paper proposes a program to estimate the hazard function in a proportional hazards model and also to treat more complex observation schemes involving general censored and left-truncated data. The hazard function estimator is defined non-parametrically as the function which maximizes a penalized likelihood, and the solution is approximated using splines. The smoothing parameter is chosen using approximate cross-validation. Confidence bands for the estimator are given. As an illustration, the age-specific incidence of dementia is estimated and one of its risk factors is studied.

Key words: Hazard estimation, proportional hazards, penalized likelihood, truncation, interval-censoring.

1 Introduction

In survival analysis, the model of choice in many applications is the Cox model [1] which allows estimation of relative risks without imposing parametric assumptions on the baseline hazard function. Using this model, both right censoring and left truncation can be handled. However, more complex observation schemes cannot be treated. For instance, in cohort studies, the subjects are observed at specific times of visits and the event of interest frequently occurs between two visits: this produces interval-censored data. Another drawback of using the Cox model is that a smooth estimate of the hazard function is not available.

Turnbull [2] proposed a non-parametric maximum likelihood estimator of the survival function for arbitrarily censored and truncated data. This method was extended [3] to the proportional hazards model in continuous time for such cases. These methods cannot be used to estimate the hazard function, which has often a meaningful interpretation in epidemiology. In particular, if age has been chosen as the time scale, the hazard function is the age-specific incidence of the disease [4]. In epidemiology, age-specific incidence is most often estimated using the person-years method. Another solution is to smooth the Nelson-Aalen estimator by kernel methods. However, neither method can accommodate interval censoring.

An alternative approach is to define the estimator non-parametrically as the function which maximizes the penalized likelihood. The solution is then approximated using splines. Such an approach has been proposed in the case of right censored data [5] and in the case of left truncated and right censored data [6].

This paper presents a computer program called PHMPL which implements this approach. PHMPL can be used to estimate regression parameters in a proportional hazards model with interval censoring and left truncation. It can also be used simply to obtain smooth estimates of the hazard function and to plot the hazard's curve, in order to check the proportional hazards assumption, for example.

In section 2 we present the model, a method for automatically choosing the smoothing parameter and another for obtaining confidence bands. In section 3, PHMPL is described and in section 4, an application is demonstrated.

2 Computational methods

This section presents a brief description of the methodology used. A more detailed presentation can be found in a previous work [7].

Let $X_1, X_2, ..., X_n$ be a sample of n positive random variables with common survival function. Thereafter, we denote by λ the hazard function and Λ the cumulative hazard function of X. The observation X_i is interval-censored if the only information known about it is that it lies in a known interval $A_i = [L_i, R_i] \subset I\!R^+$; $L_i \leq X_i \leq R_i$. Right-censoring is just a particular case of interval-censoring with $R_i = +\infty$. If $L_i = R_i$ $(= X_i)$ then X_i is uncensored. X_i is left-truncated if it is observed conditionally on the event $X_i > \mathcal{L}_i$, where $\mathcal{L}_i$ is the truncating time.

In general, under the usual assumptions regarding censoring and truncat-

ing mechanisms, the log-likelihood can be written:

$$l(\lambda) = \sum_{i=1}^{n} \log\left(\frac{e^{-\Lambda(L_i)} - e^{-\Lambda(R_i)}}{e^{-\Lambda(\mathcal{L}_i)}}\right), \qquad \mathcal{L}_i < L_i < R_i. \tag{1}$$

If $L_i = R_i$ (uncensored observation) the numerator is $\lambda(L_i)e^{-\Lambda(L_i)}$.

2.1 Penalized likelihood

Most often, the hazard function can be expected to be smooth. A possible means for introducing such *a priori* knowledge is to penalize the likelihood by a term which takes large values for rough functions.

We define the penalized log-likelihood as:

$$pl(\lambda) = l(\lambda) - \kappa \int \lambda''^2(u)du \tag{2}$$

where l is the log-likelihood defined previously in (1) and κ is the smoothing parameter which must be positive; κ controls the balance between the fit to the data and the smoothness of the function. Maximization of (2) in the desired class of function defines the maximum penalized likelihood estimator (MPnLE) $\hat{\lambda}$.

2.2 Approximation using splines

The MPnLE cannot be calculated explicitly, but can be approximated using splines. We use M-splines, which are a variant of B-splines, and I-splines which are integrated M-splines [8].

A spline function is completely defined by a sequence of increasing knots $(t_1, ..., t_l)$ and the coefficients $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T$ of the splines. In our approximation we use splines of order 4 (also called cubic splines). Therefore, we

have $m = l + 2$ parameters to estimate the hazard function. In the program, a knot is set on the first and last data points and the other knots are put equidistantly between them. Theoretically, the more knots, the better the approximation. Increasing the number of knots does not deteriorate the MPnLE: this is because the degree of smoothing in the penalized likelihood method is tuned by the smoothing parameter κ and not by the number of splines. On the other hand, once a sufficient number of knots is established, there is no advantage in adding more. Moreover, the more knots, the longer the running time, especially if there is a search for the smoothing parameter; some numerical problem can arise, particularly for a large number of knots. That is why the maximum number of knots is limited to 25. So it is recommended to start with a small number of knots (e.g. 7) and increase the number of knots until the graph of the hazard function remains unchanged (rarely more than 12 knots).

Since M-splines are nonnegative and I-splines are monotonically increasing, the monotonicity constraint for a function represented on a basis of I-splines can be fulfilled by constraining the coefficients to be positive. Thus the estimator $\hat{\lambda}(.)$ is approximated by a linear combination of m M-splines $\tilde{\lambda}(.) = \sum_{j=1}^{m} \theta_j M_j(.)$. With the same vector of coefficients $\boldsymbol{\theta}$, we get the cumulative hazard function with I-splines and the hazard function with M-splines.

The approximation $\tilde{\lambda}$ of $\hat{\lambda}$ is the function belonging to the space generated by the basis of splines which maximizes $pl(\lambda)$. The estimated vector $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ for a fixed κ is obtained by maximizing the log-likelihood using the Marquardt algorithm [9], which is a compromise between the steepest descent and the

Newton-Raphson algorithms.

2.3 Selection of the smoothing parameter

We can provide an empirical estimate of the smoothing parameter κ or use an automatic choice by using an approximation of the standard cross-validation score [5].

For an automatic choice, we maximize the following approximate cross-validation score:

$$\overline{V}(\kappa) \simeq l(\hat{\boldsymbol{\theta}}) - trace\left(\left[\hat{\boldsymbol{H}} - 2\kappa\boldsymbol{\Omega}\right]^{-1}\hat{\boldsymbol{H}}\right) \quad (3)$$

where $\hat{\boldsymbol{H}}$ is the converged Hessian $\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{\Omega}$ is the penalized part of the converged Hessian. We can interpret $trace\left(\left[\hat{\boldsymbol{H}} - 2\kappa\boldsymbol{\Omega}\right]^{-1}\hat{\boldsymbol{H}}\right)$ as the model degrees of freedom (mdf). Indeed mdf decrease in κ from m (if $\kappa = 0$) to 2 (if $\kappa \to +\infty$) which is the number of degrees of freedom of a straight line.

To maximize the approximate cross-validation score (3), we use a golden section search. In some cases, the search for the smoothing parameter may not be reliable because of local extrema. Thus, the estimate of the smoothing parameter is not optimal. This can be examined by taking different starting points. Moreover, it seems that the cross-validation score tends to under-smoothe, especially for small samples, so the smoothing parameter may be fixed *a priori* in the program. For very small samples with very few events, the estimated hazard function is so dependent on the smoothing parameter that there is no point in plotting it.

2.4 Approximate Bayesian confidence bands

We use a Bayesian technique to generate confidence bands for penalized likelihood estimators [5], [10]. After a Gaussian approximation, the covariance matrix of $\boldsymbol{\theta}$ is $-\left[\frac{1}{2}\hat{\boldsymbol{H}} - \kappa\boldsymbol{\Omega}\right]^{-1}$. Therefore, an approximate 95% Bayesian confidence interval for $\tilde{\lambda}$ at point x is: $\tilde{\lambda}(x) \pm 1,96\tilde{\sigma}(x)$, where the approximate standard error is:

$$\tilde{\sigma}(x) = \sqrt{\boldsymbol{M}(x)^T \left[-\frac{1}{2}\hat{\boldsymbol{H}} + \kappa\boldsymbol{\Omega}\right]^{-1} \boldsymbol{M}(x)},$$

where $\boldsymbol{M}(x) = (M_1(x), ..., M_m(x))^T$. To obtain approximate Bayesian confidence bands for $\tilde{\Lambda}$, hence for the survival function, we use the same formula with the I-spline basis.

2.5 Regression model

The penalized likelihood can be applied for estimating the hazard function in a proportional hazards regression model defined by: $\lambda_i(.) = \lambda_0(.)\exp(\boldsymbol{z}_i\boldsymbol{\beta})$, where $\lambda_0(.)$ is the baseline hazard function, $\boldsymbol{\beta}$ is a vector of regression parameters and $\boldsymbol{z}_i$ the vector of covariates for subject i. The penalized log-likelihood used is therefore:

$$l[\lambda_0(.), \boldsymbol{Z}\boldsymbol{\beta}] - \kappa \int \lambda_0^{''2}(u)du \tag{4}$$

where $\boldsymbol{Z}$ is the matrix with rows equal to $\boldsymbol{z}_i$, $i = 1, ..., n$.

In our method the selection of the smoothing parameter takes the most time. When covariates are included, a two-step search is used: firstly function (2) is maximized, ignoring the explanatory variables, to obtain an estimator of κ and a good initial guess of $\tilde{\lambda}_0(.)$; then, function (4) is maximized with

κ fixed to obtain $\hat{\boldsymbol{\beta}}$ and $\tilde{\lambda}_0(.)$. The regression parameters and the baseline function are estimated simultaneously by the Marquardt algorithm. Time-varying covariates cannot be handled with PHMPL.

3 Computer program

The computer program PHMPL was written in FORTRAN 77. The program runs on a Unix station or on a Personal Computer. No external functions or subroutines are needed. The running time remains acceptable even for large samples.

3.1 Input

The program requires two different input files: the data file and the parameter file.

The data file contains as many lines as subjects. The first three columns of the data file define the times of truncation and censorship. Missing values are not allowed in the first three columns. The first column is the time at entry into the study ($\mathcal{L}_i$) and is equal to 0 if there is no left-truncation. The second and the third columns are the left and the right boundaries of the interval in which the outcome occurred (L_i, R_i). If a subject did not experience the outcome, in case of right-censoring, the second column will be the time of censorship and the third must be coded -1. If the subject is left censored, the second column will be 0 and the third column will be the time of censorship. If the time of outcome is known precisely, the same time has to be given to the second and third columns.

The subsequent columns contain the value of the explanatory variables. The missing value in the explanatory variables must be coded -32768. If a specific variable is included in the model, subjects with a missing value for this variable are excluded from the analysis.

Each column must be separated by one or several spaces.

The parameter file, PHMPL.INF, contains the parameters of the analysis. The following information has to be given:

- the name of the data file
- the number of subjects
- the number of explanatory variables included in the data file
- the name of the explanatory variables and an indicator that shows whether the variable is included in the model (1 = included and 0 = not included)
- the number of knots (from 5 to 25)
- the indication whether to search the smoothing parameter automatically or not (0 = automatically and 1 = fixed)
- the initial value of the smoothing parameter
- the indication to record the hazard function and the survival function (1 = saved and 0 = not saved)
- the name of these files

An example of PHMPL.INF file is given in appendix 1.

3.2 Output

The log-likelihood, the value of the estimated smoothing parameter, and the model degrees of freedom are given on the screen.

The file REGR.RES is created automatically and contains the coefficient estimates of each explanatory variable. This file is not created if explanatory variables are not included in the model. This file contains the value of the log-likelihood, the number of regression parameters, the number of subjects and the number of events. For each variable, its name, the value of the coefficient and its standard error, the value of the Wald test, the value of the relative risk and its confidence interval are given.

At the user's request, two files are created whose names have to be written in the parameter file. The first contains the coordinates for plotting the hazard function and its confidence bands between the first and the last knots; the second contains the coordinates for plotting the survival function and its confidence bands. Note that if explanatory variables are included, the functions saved are the baseline functions.

4 Application

To illustrate the use of PHMPL we present an application for modeling the risk of developing dementia. The application is based on the Paquid research program [11], a prospective cohort study on mental and physical aging that evaluates social environment and health status. The target population consists of subjects aged 65 years and older living at home in southwestern

France. The baseline variables recorded included socio-demographic factors, medical history and psychometric tests. Subjects were re-evaluated 1, 3 and 5 years after the initial visit. Age was used as the basic time scale in order to obtain the age-specific incidence of dementia. Prevalent cases were excluded from the sample, so this produced a left-truncation problem. The age of onset of dementia was left-truncated by the age of the subject at inclusion in the study, and was right censored if the subject had not developed dementia at the time last seen. The sample consisted of 2881 subjects and during the 5 years of follow-up, 190 incident cases of dementia were observed. The information available on incident cases of dementia was the date of the most recent visit without dementia and the date of the first visit with dementia. Thus the age of onset of dementia for these 190 subjects was interval-censored. The program allows the treatment of these data which are both interval-censored and left-truncated.

One explanatory variable was considered: primary school diploma. In the sample 932 subjects did not possess the primary school diploma while 1949 did. The PHMPL.INF file used for analyzing the model is given in appendix 1.

The estimate of the relative risk to develop dementia between subjects with the diploma and those without was 1.93 (95 % confidence interval $[1.44, 2.57]$). To verify the proportional hazards assumption for primary school diploma, we did two separate analyses. The non-parametric estimates of the risks displayed in Figure 1 confirm that subjects without the diploma have a higher risk of dementia; however, the hazards do not seem to be proportional. We therefore suggest: $\lambda(t) = \lambda_0(t + \beta z)$ where z is a covariate

taking values 0 or 1 according to whether the subjects have the diploma. Thus, the program allows to check whether the proportional hazards model is appropriate by comparing the two hazard curves obtained through two separate analyses; when the proportional hazards assumption does not hold, it is necessary to perform separate analyses for each group, as shown in this example. Another solution, not feasible with PHMPL, is to perform a stratified analysis [12] allowing different baseline hazards for the strata but the same effects for other covariates.

5 Availability

The portable code implementing the algorithm is available to the public at no charge at http://www.isped.u-bordeaux2.fr .

References

[1] D.R. Cox, Regression models and life tables (with Discussion), Journal of the Royal Statistical Society, Series B 34 (1972) 187-220.

[2] B. W. Turnbull, The empirical distribution function with arbitrarily grouped, censored and truncated data, Journal of the Royal Statistical Society, Series B 38 (1976) 290-295.

[3] A. Alioum and D. Commenges, A proportional hazards model for arbitrarily censored and truncated data, Biometrics 52 (1996) 512-524.

[4] D. Commenges, L. Letenneur, P. Joly, A. Alioum and J.-F. Dartigues, Modelling age-specific risk: application to dementia, Statistics in Medicine 17 (1998) 1973-1988.

[5] F. O'Sullivan, Fast computation of fully automated log-density and log-hazard estimators, SIAM Journal on Scientific and Statistical Computing 9 (1988) 363-379.

[6] C. Gu, Penalized likelihood hazard estimation: a general procedure, Statistica Sinica 6 (1996) 861-876.

[7] P. Joly, D. Commenges and L. Letenneur, A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia, Biometrics 54 (1998) 185-194.

[8] J. O. Ramsay, Monotone regression splines in action, Statistical Science 3 (1988) 425-461.

[9] D. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, SIAM Journal of Applied Mathematics 11 (1963) 431-441.

[10] G. Wahba, Bayesian "confidence intervals" for the cross-validated smoothing spline, Journal of the Royal Statistical Society, Series B 45 (1983) 133-150.

[11] L. Letenneur, D. Commenges, J.-F. Dartigues and P. Barberger-Gateau, Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France, International Journal of Epidemiology 23 (1994) 1256-1261.

[12] J. P. Klein and M. L. Moeschberger, Survival Analysis Techniques for censored and truncated data, Statistical Science, Chap. 9 (Springer, 1997).

Appendix 1

Example of PHMPL.INF file used for regression analysis:

```
dementia
2881
1
diploma 1
12
0
1.e6
1
surv.dat
hazard.dat
```

The main program (PHMPL) reads the data in the file "dementia" that contains 2881 lines (therefore, 2881 subjects). One explanatory variable is stored in the file (diploma) and is included in the proportional hazards model. There are 12 knots and the smoothing parameter is estimated automatically. The initial value of the smoothing parameter is 10^6. The coordinate of the baseline survival function and its confidence intervals are saved in the file surv.dat. The baseline risk and its confidence intervals are saved in the file hazard.dat.

The result of the regression analysis is in the file REGR.RES:

log-likelihood : -886.2611

number of parameters : 1 number of subjects : 2881 number of events :

190

Variable : diploma

beta : 0.6553 SE(beta) : 0.1468

Wald : 4.4627

RR : 1.9258 95% IC : 1.4441 2.5681

Captions for Figures

Figure 1: Approximation of the hazard function of dementia for subjects without primary school diploma (dotted line) and for those with it (solid line).