Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Fitting regression models to interval-censored observations of illness-death models

Célia Touraine
University of Bordeaux

Thomas A. Gerds
University of Copenhagen

Pierre Joly
University of Bordeaux

Abstract

The irreversible illness-death model allows subjects to move from an initial state ("health state") to a terminal state ("death state") either directly or through an intermediate state ("disease state"). Disease onset times may not be known exactly, for example if the disease status of a patient can only be assessed at regular visits. In this situation the disease onset times are usually interval-censored. This article presents the **SmoothHazard** package for R. It implements algorithms for simultaneously fitting regression models to the three transition intensities of an illness-death model where the times to the intermediate state may be interval-censored data. The three baseline hazard functions are either parametrized according to Weibull distributions or approximated by M-splines. For a given set of covariates, the intensities models can be combined into predictions of cumulative event probabilities and life expectancies.

Keywords: illness-death model, interval-censored data, left-truncated data, survival model, proportional regression models, Smooth Transition intensities, Weibull.

1. Introduction

The irreversible illness-death model is a special multi-state model which has many applications for example in medical research. The model allows subjects to make transitions from an initial state (health) to a terminal state (death) either directly or via an intermediate state (disease), see Figure 1. If the exact transition times are observed, standard procedures can be used to estimate regression models for the three transitions (de Wreede *et al.* 2011). In particular, the regression coefficients can be estimated using Cox partial likelihood without the need to specify the baseline intensities. However, this possibility is lost when the transition times from the initial state to the transient state are interval censored. For example, it may not be possible to determine the exact onset time of dementia for a patient diagnosed at time R. Instead it is only known that the patient was last seen dementia-free at time L. Thus,

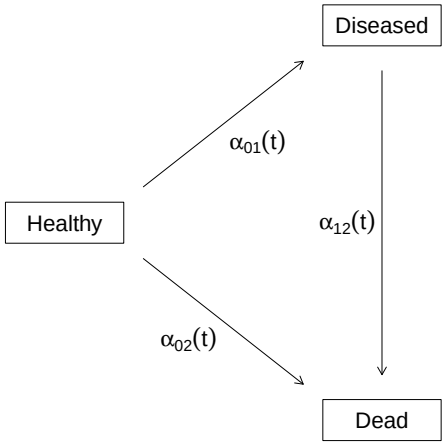


Figure 1: The irreversible illness-death model

the onset time is interval censored between L and R. The algorithms of the **SmoothHazard** package implement methods for estimating regression models under this type of censoring if the transition times into the absorbing state (e.g. death of the patient) are either known exactly or right censored (Joly *et al.* 2002).

Implemented are a parametric and a semi-parametric estimation approach. For the parametric approach, the Weibull distribution is used for the baseline transition intensities and the parameters are estimated by maximising the likelihood. For the semi-parametric approach, M-splines are used to approximate the baseline transition intensities and the parameters are estimated maximising a penalized likelihood. The methods allow delayed entry of the subjects, i.e. that the event times are left truncated.

Section 2 presents the model and the likelihood. Section 3 presents the estimation methods. Section 4 provides some examples illustrating **SmoothHazard** functions.

2. Data, model and questions

In order to illustrate the functionality of the package we provide a random subset containing data from 1000 subjects that were enrolled in the Paquid study (Letenneur *et al.* 1999), a large cohort study on mental and physical aging.

```
1 data(Paq1000)
```

The population consists of subjects aged 65 years and older living in Southwestern France. The event of interest is dementia and death without dementia is a competing risk. Furthermore, the time to dementia onset is interval censored.

In this subset `186` subjects are diagnosed as demented and `724` died from whom `597` without being diagnosed as demented before. Because of interval censoring more than `186` should have been demented, more than (respectively more than `127` should have been dead with dementia and less than `597` should have been dead with dementia (see Figure 2). There are two covariates in this subset: gender (`578` women and `422` men) and primary school diploma (`762` with diploma and `238` without diploma). Age is chosen as the basic time scale

and subjects are non demented at entry into stydy. Consequently, we need to deal with left-truncated event times.

```
head(Paq1000)
```

```
  dementia death      t0       l        r        t certif gender
1        1     1 72.3333 82.34014 84.73303 87.93155      0      0
2        0     1 77.9167 78.93240 78.93240 79.60048      0      1
3        0     1 79.9167 79.91670 79.91670 80.92423      0      0
4        0     1 74.6667 78.64750 78.64750 82.93501      1      1
5        0     1 76.6667 76.66670 76.66670 79.16636      0      1
6        0     0 66.2500 71.38070 71.38070 84.16975      1      0
```

Each row in the data corresponds to one subject. The variables `dementia` and `death` are the status variables (1 for an event, 0 for right censoring) for events dementia and death, respectively. The variable `t0` contains ages of subjects at entry. The variables `l` and `r` contain the left and right endpoints of the censoring intervals. For demented subjects, `r` is the age at the diagnostic visit and `l` is the age at the previous one. For non demented subjects, `l` and `r` are the age at the latest visit without dementia (`l`=`r`). The variable `t` is the age at death or at latest news. `certif` and `gender` are binary covariates.

The function `idm` computes estimates for the three transition intensities α_{01}, α_{02}, α_{12} which are age-specific incidence of dementia, age-specific mortality rate of dementia-free subjects and age-specific mortality rate of demented sujects, respectively. Proportional transition intensities regression models allow for covariates :

$$\alpha_{hl}(t|Z_{hli}) = \alpha_{0hl}(t)\exp\{\beta_{hl}^T Z_{hli}\}; \quad hl \in \{01, 02, 12\}$$

where α_{0hl} are baseline transition intensities, Z_{hli} are covariates vectors and β_{hl} are vectors of associated regression parameters.

Covariates must be specified at the right side of the `formula01`, `formula02` and `formula12` arguments of the `idm` function (`~1` for no covariates).

Interval censoring and left truncation must be specified at the left side of the formula arguments using the `Hist` function. For left-truncated data, the `entry` argument of `Hist` must contain the vector of delayed entry times. For interval-censored data, the `time` argument of `Hist` must contain a list of the left and right endpoints of the intervals.

A call of the `idm` function looks as follows:

```
fit <- idm(formula01=Hist(time=list(l,r),event=dementia,entry=t0)~certif,
           formula02=Hist(time=t,event=death,entry=t0)~certif+gender,
           formula12= ~ 1,
           data=Paq1000)
```

where the `data` argument contains the data frame in which to interpret the variables of `formula01`, `formula02` and `formula12`.

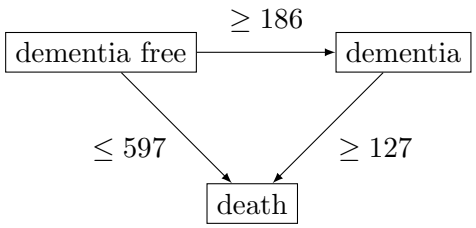


Figure 2: The exact number of transitions in the illness-death model with interval-censored time to disease is unknown.

Note that the left side of `formula12` does not need to be filled because all the data informations are already contained in `formula01` and `formula02`. In fact, the `formula12` argument is required only if we want the covariates impacting transition 12 different from those impacting transition 02.

Questions ? TODO

3. Fitting the illness-death model based on interval-censored data

The `idm` function computes estimates for the three transition intensities:

$$\alpha_{hl}(t|Z_{hli}) = \alpha_{0hl}(t)\exp\{\beta_{hl}^T Z_{hli}\}; \quad hl \in \{01, 02, 12\}$$

In the situation where time to disease and time to death are not interval censored the regression coefficients can be estimated by the partial likelihood method (Cox 1975) without the need to specify or estimate the baseline hazard functions $\alpha_{0hl}(t)$. For interval-censored transition times to state 1 the situation is more complex. It turns out that we have to estimate all parameters simultaneously and that we need a model for the baseline transition intensity functions. This can be seen by inspecting the likelihood function.

For subject i, let us denote the conditional event-free survival function by

$$S(t|Z_{01i}, Z_{02i}) = e^{-A_{01}(t|Z_{01i}) - A_{02}(t|Z_{02i})}$$

where $A_{hl}(.|Z_{hli})$ are the conditional cumulative intensity functions:

$$A_{hl}(t|Z_{hli}) = \int_0^t \alpha_{hl}(u|Z_{hli})du$$

.

We set $\delta_{1i} = 1$ ($\delta_{1i} = 0$) if subject i has (has not) been observed diseased, and $\delta_{2i} = 1$ ($\delta_{2i} = 0$) if subject i is (is not) dead.

If $\delta_{2i} = 0$, T_i is time to death; if $\delta_{2i} = 0$, death event is right-censored at T_i. We denote by L_i and R_i the interval censoring times. If subject i has been observed diseased at time R_i and has last been seen non diseased at time L_i ($L_i < R_i$), time to disease is interval-censored between L_i and R_i. The likelihood contribution for subject i is:

$$\mathcal{L}_i = \frac{1}{S(T_{0i}|Z_{01i}, Z_{02i})} \int_{L_i}^{R_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i}) \frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}} \big(\alpha_{12}(T_i|Z_{12i})\big)^{\delta_{2i}} du \quad (1)$$

If subject i has never been seen diseased, time to disease is right-censored and the interval censoring times are set to the right censoring time ($L_i = R_i$). The likelihood contribution for subject i is:

$$\mathcal{L}_i = \frac{1}{S(T_{0i}|Z_{01i}, Z_{02i})} \left(S(T_i|Z_{01i}, Z_{02i}) \big(\alpha_{02}(T_i|Z_{02i})\big)^{\delta_{2i}} + \int_{L_i}^{T_i} S(u|Z_{01i}, Z_{02i}) \alpha_{01}(u|Z_{01i}) \frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}} \big(\alpha_{12}(T_i|Z_{12i})\big)^{\delta_{2i}} du \right) \quad (2)$$

If time to disease and time to death are both right-censored at the same time, we have $L_i = R_i = T_i$ and the integral value in (2) is zero.

3.1. The Weibull parametrization

The default estimation method in function `idm` computes the maximum likelihood estimate for the three transition intensities using a Weibull parametrization for the baseline transition intensities:

$$\alpha_{0hl}(t) = a_{hl}\ b_{hl}^{a_{hl}}\ t^{a_{hl}-1}; \quad hl \in \{01, 02, 12\}.$$

where a_{hl} and b_{hl} are shape and scale parameters.

```
fit.weib <- idm(formula02=Hist(time=t,event=death,entry=t0)~certif+gender,
                formula01=Hist(time=list(l,r),event=dementia,entry=t0)~certif+gender,
                data=Paq1000,intensities="Weib")
print(fit.weib)
```

```
 Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = t0) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = t0) ~ certif + gender, data = Paq1000, intensities = "Weib")

Illness-death Model using a parametric approach with a Weibull distribution for the intens

number of subjects:  1000
number of events '0-->1':  186
number of events '0-->2' or '0-->1-->2':  724
number of covariates:  2 2 2

             coef SE.coef     HR          CI       Wald  p.value
certif_01 -0.5194  0.2016 0.5949 [0.40;0.88]  6.6403066 0.009970
gender_01 -0.1221  0.1598 0.8851 [0.65;1.21]  0.5835582 0.444921
certif_02  0.1268  0.1263 1.1352 [0.89;1.45]  1.0073060 0.315549
gender_02  0.5363  0.1200 1.7096 [1.35;2.16] 19.9881686  < 1e-04
certif_12 -0.2079  0.2322 0.8123 [0.52;1.28]  0.8017104 0.370582
gender_12  0.5792  0.1865 1.7846 [1.24;2.57]  9.6480046 0.001896
```

```
               Without cov  With cov
Log likelihood   -3075.308 -3048.791

Parameters of the Weibull distribution: 'S(t) = exp(-(b*t)^a)'
      alpha01    alpha02    alpha12
a 11.18802123 8.62750229 7.50199862
b  0.01099806 0.01078284 0.01294115

----
Model converged.
number of iterations:  8
convergence criteria: parameters= 1.1e-07
                    : likelihood= 6.4e-07
                    : second derivatives= 4.4e-10
```

Maximization algorithm

The vectors of parameters for the baseline transition intensities a_{hl} and b_{hl} and the vectors of regression parameter $\hat{\beta}_{hl}$ are obtained simultaneously by maximizing the log-likelihood using a combination of a Marquardt's algorithm (Marquardt 1963) which is a robust Newton-like algorithm and a steepest descent algorithm. Using the Marquardt's algorithm, few iterations are needed if the initial value is judiciously chosen. The Marquardt's algorithm step involves a line search with a step reduction if the new point is not better. Using the steepest descent algorithm, the convergence is slower. The steepest descent step involves a full line search and is attempted only if the Marquardt's algorithm step has failed, due generally to a difficulty to inverse the Hessian matrix of the log-likelihood. We stop the iterations when the differences between two consecutive parameters values, log-likelihood values, and gradient values is small enough. The default convergence criteria are 10^{-5}, 10^{-5} and 10^{-3} and can be changed by means of the `eps` argument. The variances of parameter estimates are estimated using the inverse of the matrix of the second derivatives at convergence.

3.2. The penalized likelihood

Another estimation method in `idm` permits to get smooth transition intensities without parametric specification. Using the option `intensities="Splines"`, a maximum penalized likelihood estimate is computed using a spline approximation for the three transition intensities α_{01}, α_{02}, α_{12}.

To force smoothness of intensity functions, we penalize the likelihood by a term relating to the curvature of the intensity functions that is the quare of the second derivates.

The penalized log-likelihood (pl) is defined as:

$$pl = l - \kappa_{01} \int \alpha_{01}^{''\,2}(u)du - \kappa_{12} \int \alpha_{12}^{''\,2}(u)du - \kappa_{02} \int \alpha_{02}^{''\,2}(u)du \tag{3}$$

where l is the log-likelihood and κ_{01}, κ_{02} and κ_{12} are three positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions.

Maximization of (3) defines the maximum penalized likelihood estimators (MPLE) $\hat{\alpha}_{01}$, $\hat{\alpha}_{02}$ and $\hat{\alpha}_{12}$.

Approximation via splines

A spline of order k is a piecewise polynomial functions of degree $k-1$. The places where the polynomial pieces connect are the knots. Associated with a knot sequence t, basis splines can be combined linearly to yield any other spline associated with t. $\hat{\alpha}_{01}$, $\hat{\alpha}_{02}$ and $\hat{\alpha}_{12}$ are approximated using linear combination of M-splines (Ramsay 1988). For $hl \in \{01, 02, 12\}$:

$$\tilde{\alpha}_{hl}(x) = \sum_{i=1}^{n} a_i M_i(x)$$

where n is the number of free parameters.
The non-negativity of $\tilde{\alpha}_{hl}$ is assured by constraining the coefficients a_i to be positive. In practice, we estimate parameters θ_i such that $a_i = \theta_i^2$ which maximize the penalized likelihood.
A M-spline of order k is computed using the following recursion:
For $k = 1$,

$$M_j(x|1,t) = \begin{cases} \frac{1}{(t_{j+1}-t_j)} & \text{if } t_j \leq x < t_{j+1} \\ 0 & \text{elsewhere} \end{cases}$$

For $k > 1$,

$$M_j(x|k,t) = \begin{cases} \frac{k\left[(x-t_j)M_j(x|k-1,t)+(t_{j+k}-x)M_{j+1}(x|k-1,t)\right]}{(k-1)(t_{j+k}-t_j)} & \text{if } t_j \leq x < t_{j+k} \\ 0 & \text{elsewhere} \end{cases}$$

where $t = t_1, ..., t_{n+k}$ is a knot sequence.
The M-spline family is particularly appealing to statisticians because each M_i has the properties of a probability density function over the interval $[t_i, t_{i+k}]$. Among them, we have $\int M_i(x)dx = 1$
One can associate to each M-spline, the integrated splines or I-splines I_i, $i = 1, \ldots, n$ such that $I_i(x|k,t) = \int_{t_k}^{x} M_i(u|k,t)du$. Given the coefficients a_i, we can approximate estimators of the cumulative transition intensities $\hat{A}_{hl}$ using a linear combination of I-splines:

$$\tilde{A}_{hl}(x) = \sum_{i=1}^{n} a_i I_i(x)$$

Because M-splines are non-negative, the positivity constraint on a_i ensures that the $\tilde{A}_{hl}$ are monotically increasing. Each M_j is piecewise polynomial of degree $k-1$ and each associated I_j is piecewise polynomial of degree k. In the package we use cubic M-splines *i.e.* $k = 4$.

Choice of the knots

The knots sequence has some properties to ensure continuity conditions. Among them, we have: $t_1 = \ldots = t_k$ and $t_{n+1} = \ldots = t_{n+k}$. The number of free parameters n corresponds to $k+$ the number of knots interior to $[t_k, t_{n+1}]$

In **SmoothHazard**, the knots are put equidistantly between them by default. The `knots` argument can be fulfilled to choose their location freely but in general the shape of a spline function is not very sensitive to knot placement. However, there must be several data points between each pair of different knots and there must be a knot before or at the first data point and after or at the last data point. Increasing the number of data points between a pair of knots leads to a better defined curve.

The number of knots can be specified in the `n.knots` argument. It must be understood as the number of different knots *i.e.* the number of knots from t_k to t_{n+1}. The default is 7 on the three transitions which leads to a number of free parameters one one transition $n = 3+5 = 9$. Increasing the number of knots in a region leads to a greater flexibility of the function in that region. The number of knots and their location can be chosen differently for each transition.

Smoothing parameters

The default values for the smoothing parameters are suitable for the avalaible `Paq1000` data set. However, these values can be expected to be very different depending on time scale and number of subjects. They can be changed into the `kappa` argument. An approximate cross-validation technique to determine the smoothing parameters is also available with the option `CV=TRUE`. In this case, the `kappa` argument contains the initial values for the smoothing parameters. We use an approximate leave-one-out score proposed by O'Sullivan (1988) for survival models and extended for multi-state models (Commenges *et al.* 2007) for which only one estimation of the model is required by tested values of the smoothing parameters.

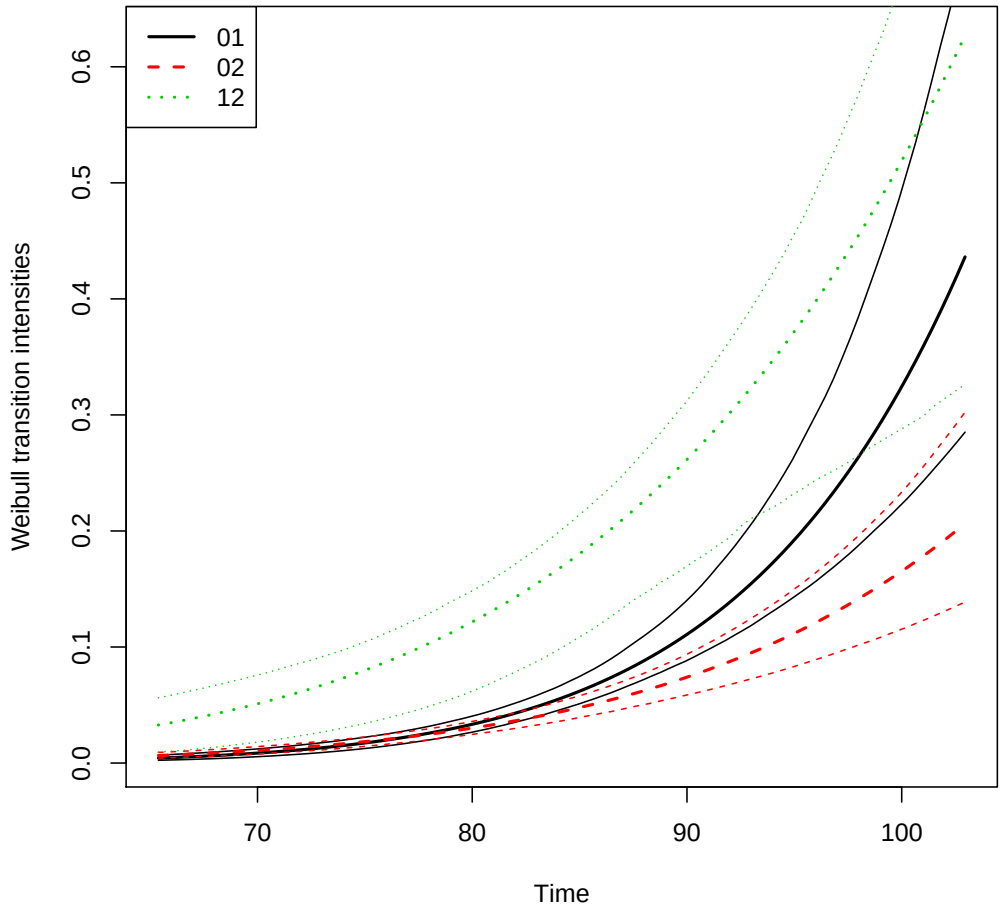
Maximization algorithm

The vectors of spline coefficients for fixed κ_{01}, κ_{12} and κ_{02} and the vectors of regression parameters $\hat{\beta}_{01}$, $\hat{\beta}_{02}$, $\hat{\beta}_{12}$ are obtained simultaneously by maximizing the penalized log-likelihood using the same maximization algorithm as with the Weibull parametrization (see Section 3.1).

Practical advices

Increasing the number of knots does not deteriorate the MPLE: this is because the degree of smoothing in the penalized likelihood method is tuned by the smoothing parameters κ_{01}, κ_{12} and κ_{02}. 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. 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. 5 or 7) and increase the number of knots until the graph of the transition intensities function remains unchanged (rarely more than 12 knots).

The choice of the smothing parameters can be fastidious. The `idm` function can be run with the approximate cross-validation option. However, the running time is very long and an empirical technique can be preferred. It consists in repeting the idm running trying different smoothing parameters. After each estimation, the transition intensities must be plotted, for example with the `plot` function. For the curves that seem over-smooth, the associated smoothing parameter must be reduced. For the curves that seem under-smooth, the associated smoothing parameter must be increased.



4. Predicting parameters of life

Most often in illness-death models, the functions of interest are the transition intensities. In our application, $\alpha_{01}(.)$, $\alpha_{02}(.)$ and $\alpha_{12}(.)$ corresponds to age-specific incidence of dementia, age-specific mortality rate of non demented subjects and age-specific mortality rate of demented subjects. However, other functions/quantities which can be expressed in terms of the transition intensities (Touraine *et al.* 2013) and may provide additional information and have a more natural interpretation.

The fonction `idm` returns an "idmWeib" or "idmSplines" class object depending on the parametrization of the transition intensities (Weibull or splines). These objects can be used in argument of the `predict.idmWeib` and `predict.idmSplines` functions in order to obtain transition probabilities and cumulative probabilities:

```
TP_fit.weib <- predict(fit.weib,s=70,t=80,Z01=c(1,0),Z02=c(1,0),Z12=c(1,0))
TP_fit.weib
```

```
 null device
           1
 $p00
[1] 0.7668752 0.6715004 0.7659254

$p01
[1] 0.06363150 0.04632911 0.10304070

$p11
[1] 0.3938491 0.3253956 0.7917419

$p12
[1] 0.6061509 0.2082581 0.6746044

$p02_0
[1] 0.1367741 0.1416909 0.2358087

$p02_1
[1] 0.032719150 0.008783744 0.044995691

$p02
[1] 0.1694933 0.1689177 0.2537382

$F01
[1] 0.09635065 0.06175630 0.14105089

$F0.
[1] 0.2331248 0.2340746 0.3284996
```

They can also be used in argument of the `lifexpect` function to obtain life expectancies:

```
LE_fit.weib <- lifexpect(fit.weib,s=90,Z01=c(1,0),Z02=c(1,0),Z12=c(1,0),CI=FALSE)
LE_fit.weib
```

```
$life.in.0.expectancy
[1] 4.54907

$life.expectancy.nondis
[1] 5.906281

$life.expectancy.dis
[1] 3.706724
```

References

Commenges D, Joly P, Gégout-Petit A, Liquet B (2007). "Choice between Semi-parametric Estimators of Markov and Non-Markov Multi-state Models from Coarsened Observations." *Scandinavian Journal of Statistics*, **34**(1), 33–52.

Cox DR (1975). "Partial Likelihood." *Biometrika*, **62**, 269–276.

de Wreede LC, Fiocco M, Putter H (2011). "mstate: An R Package for the Analysis of Competing Risks and Multi-State Models." *Journal of Statistical Software*, **38**(7), 1–30. URL http://www.jstatsoft.org/v38/i07.

Joly P, Commenges D, Helmer C, Letenneur L (2002). "A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia." *Biostatistics*, **3**(3), 433–443.

Letenneur L, Gilleron V, Commenges D, Helmer C, Orgogozo J, Dartigues J (1999). "Are sex and educational level independent predictors of dementia and AlzheimerâĂŹs disease? Incidence data from the PAQUID project." *Journal of Neurology, Neurosurgery & Psychiatry*, **66**(2), 177–183.

Marquardt DW (1963). "An algorithm for least-squares estimation of nonlinear parameters." *Journal of the Society for Industrial & Applied Mathematics*, **11**(3), 431–441.

O'Sullivan F (1988). "Fast computation of fully automated log-density and log-hazard estimators." *Journal on Scientific and Statistical Computing*, **9**(2), 363–379.

Ramsay JO (1988). "Monotone regression splines in action." *Statistical Science*, **3**(4), 425–441.

Touraine C, Helmer C, Joly P (2013). "Predictions in an illness-death model." *Statistical methods in medical research.*

Affiliation:

Célia Touraine
Univ. Bordeaux
ISPED
Centre INSERM U-897-Epidemiologie-Biostatistique
Bordeaux F-33000
France
E-mail: celia.touraine@isped.u-bordeaux2.fr
URL: http://www.isped.u-bordeaux2.fr/