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 follow-up 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 transition intensities estimates can be combined into predictions of transition probabilities, 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, penalized likelihood, M-splines.

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).

If the exact transition times are observed, standard procedures like those implemented in the **mstate** package can be used to estimate cumulative transition intensities, regression coefficients and transition probabilities. (de Wreede *et al.* 2011). In particular, the regression coefficients can be estimated using Cox partial likelihood (Cox 1975) without the need to specify the baseline intensities. However, this possibility is lost when the transition times

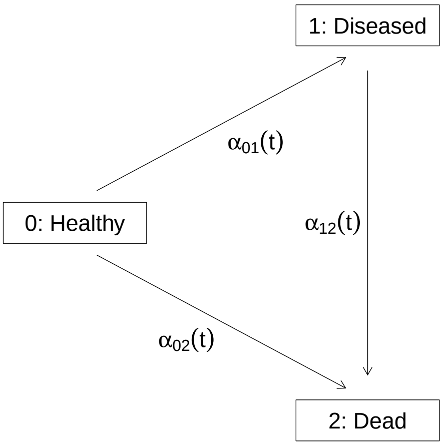


Figure 1: The irreversible illness-death model

from the initial state to the intermediate state are interval censored. For example, it may not be possible to determine the exact onset time of disease for a subject diagnosed at time R. Instead it is only known that the subject was last seen disease-free at time L. Thus, the onset time is interval censored between L and R. Furthermore, for a subject who died without being diagnosed as diseased before, it may not be possible to determine if the subject developed disease between the last time he was seen and the time of death. A usual circumvention technique to handle these subjects consists in right-censoring dead subjects without disease diagnostic at the last time they were seen without disease. By imputing onset time of disease for diseased subjects (for example in the middle of the interval censoring interval $[L, R]$), classique multistate analyses as implemented in **mstate** may be applied. However, this approach can lead to an underestimation of the transition intensity of disease (corresponding to the hazard function in survival models settings) (Joly *et al.* 2002) and biases in the regression coefficients estimates (Leffondré *et al.* 2013), especially if risk of death for diseased subjects is higher than risk of death for disease-free subjects.

The **msm** package (Jackson 2011) allows to fit Markov multi-state models to panel data (where states of each subjects are only known at a finite series of times) and could be used to fit illness-death models to data with interval-censored disease times and exact death times. But in this package, the likelihood is calculated using the Kolmogorov differential equations that relate the transition probabilities and the transition intensities. It implies a time-homogeneity assumption: constant or piecewise-constant transition intensities. The **msm** package allows for very general multi-state models where the number of possible pathways taken by a subject between two observation times can be infinite. By comparison, the illness-death model for interval-censored data is very simple with a maximum of two possible pathways between two observation times. The direct expression of the transition probabilities in terms of transition intensities can be derived making possible to estimate more flexible transition intensities without time-homogeneity assumption.

The **SmoothHazard** package allows to estimate transition intensities of a non-homogeneous Markov illness-death model under interval censoring if the transition times into the absorbing state (e.g. death of the subject) are either known exactly or right censored. Proportional transition intensities regression models allow for covariates on the three transitions. Implemented are a parametric and a semi-parametric estimation method. For the parametric approach, the Weibull distribution is used for the baseline transition intensities and the parameters (regression coefficients and Weibull parameters) are estimated by maximising the likelihood. For the semi-parametric approach, a M-splines basis is used to approximate the baseline transition intensities and the parameters (regression coefficients and spline coefficients) are estimated maximising a penalized likelihood. The methods allow delayed entry of the subjects, i.e. that the event times are left truncated.

The main functions of **SmoothHazard** are

- `idm` : for fitting illness-death model based on possibly interval-censored disease times and exact death times;
- `shr` : for fitting survival model based on possibly interval-censored event times.

Other packages can provide estimates in survival models with exact (e.g. **survival**) or interval-censored (e.g. **intcox**) event times, and estimates in illness-death model with exact transition times (**mstate**). We focus in this paper on what makes **SmoothHazard** especially appealing: being able to fit illness-death model to interval-censored data.

The above `R` functions are essentially an interface between the user and FORTRAN programs which constitute the most part of the package. An object as returned by `shr` or `idm` can then be used as argument in other `R` functions. For example, a fitted illness-death model as produced by `idm` can be used to calculate predictions for a given set of covariates: transition probabilities, cumulative event probabilities, life expectancies.

Section 2 presents the model and the likelihood. Section 3 presents the estimation methods. Section 4 briefly presents predictions that can be made in an illness-death model. Section 5 provides some examples illustrating **SmoothHazard**.

2. Model and likelihood

We consider an illness-death process $X = (X(t), t \geq 0)$. $X(t)$ has values in $\{0, 1, 2\}$. Subjects are initially disease-free ($X(0) = 0$) and may become diseased (transition $0 \to 1$) and die (transition $1 \to 2$), or die directly (transition $0 \to 2$.) X is assumed to be a non-homogeneous Markov process which means that the future evolution of the process $\{X(t), t > s\}$ depends on the current time s and only on the current state $X(s)$. X is fully characterized by the transition probabilities :

$$p_{hl}(s, t) = \mathbb{P}(X(t) = l|X(s) = h)$$

or the transition intensities which are instantaneous transition probabilities :

$$\alpha_{hl}(t) = \frac{p_{hl}(t, t + \Delta t)}{\Delta t}$$

The transition intensities in multi-state models are similar to hazard functions in survival models (see Figure 1).

We introduce covariates on each transition through proportional transition intensities regression models which are a natural extension of the Cox proportional hazard model :

$$\alpha_{hl}(t|Z_{hli}) = \alpha_{0,hl}(t) \exp\{\beta_{hl}^T Z_{hli}\}; \qquad hl \in \{01, 02, 12\} \tag{1}$$

where $\alpha_{0,hl}$ are baseline transition intensities, Z_{hli} are covariates vectors for subject i and β_{hl} are vectors of regression parameters.

In the situation where time to disease and time to death are not interval censored the regression coefficients could be estimated by the partial likelihood method without the need to specify or estimate the baseline hazard functions $\alpha_{0,hl}(t)$. 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:

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

Note that if subject i has reached state 1, the conditional survival function in state 1 between times s and t is:

$$\frac{e^{-A_{12}(t|Z_{12i})}}{e^{-A_{12}(s|Z_{12i})}}$$

We set $\delta_{2i} = 1$ ($\delta_{2i} = 0$) 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. Let us detail the likelihood contribution for subject i by distinguishing if subject i has been observed in state 1 (diseased) or not.

- If subject i has first been observed diseased at time R_i and has last been seen disease-free at time L_i ($L_i < R_i$), disease time is interval-censored between L_i and R_i. The likelihood contribution for subject i is:

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

Indeed, subject i has survived in state 0 until some time u between L_i and R_i and moved to state 1 at time u. Then, he has survived in state 1 between times u and time T_i and died at time T_i if $\delta_{2i} = 1$.

- If subject i has never been seen diseased, let us denote the last time he has been observed disease-free by R_i. The likelihood contribution for subject i is:

$$\mathcal{L}_i = S(T_i|Z_{01i}, Z_{02i})\big(\alpha_{02}(T_i|Z_{02i})\big)^{\delta_{2i}} + \int_{R_i}^{T_i} S(u|Z_{01i}, Z_{02i})\alpha_{01}(u|Z_{01i})\frac{e^{-A_{12}(T_i|Z_{12i})}}{e^{-A_{12}(u|Z_{12i})}}\big(\alpha_{12}(T_i|Z_{12i})\big)^{\delta_{2i}}du \qquad (3)$$

Indeed, if subject i has not died at T_i ($\delta_{2\mathrm{i}}$=0), he may have survived in state 0 (term at the left side of the plus sign) or he may have becomed diseased between R_i and T_i (term at the right side of the plus sign); if subject i has died at T_i, he may have moved directly from state

0 to state 2 (term at the right side of the plus sign) or he may have became diseased at some time between R_i and L_i and then died (term at the right side of the plus sign).

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 (3) is zero.

Suppose now that data are left truncated, i.e. that subjects are under observation starting from time $T_0 > 0$. Let us denote T_{0i} the time from which subject i is under observation. The left truncation condition $X(T_{0i}) = 0$ (subject i has survived in state 0 until time T_{0i}) is taken into account by dividing the above likelihood contributions by the term $S(T_{0i}|Z_{01i}, Z_{02i})$.

3. Estimation

The idm function computes estimates for the three baseline transition intensities and for the regression parameters using likelihood-based estimation methods. In the parametric method and in the semi-parametric method, respectively the likelihood and the penalized likelihood are maximized using the Levenberg-Marquardt's algorithm (Levenberg 1944; Marquardt 1963) which is a combination of a Newton-Raphson algorithm and a gradient descent algorithm (also known as the steepest descent algorithm). This algorithm has the avantage of being more robust than the Newton-Raphson algorithm while preserving its fast convergence property.

3.1. Parametric estimation

In the default estimation method of function idm, a Weibull parametrization for the baseline transition intensities is assumed:

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

where a_{hl} and b_{hl} are shape and scale parameters. The Weibull parameters a_{hl} and b_{hl} and the vectors of regression parameter $\hat{\beta}_{hl}$ are obtained simultaneously by maximizing the log-likelihood.

Confidence intervals for the regression parameters are obtained using estimated standard errors estimated by inverting the Hessian matrix of the log-likelihood. Confidence bands for the baseline transition intensities are obtained using a simulation-based approach explained below (section 4.1). (PIERRE: is it really right ? -> to be checked in the fortran programs of the package ?)

3.2. Semi-parametric estimation

The other estimation method in the function idm permits to relax the strict parametric assumptions of the Weibull regression models: linear combinations of M-splines are used to approximate the three baseline transition intensities and the maximization of a penalized likelihood ensures to obtained smooth estimates of them.

The penalized likelihood approach

To control the smoothness of the estimated intensity functions, we penalize the log-likelihood by a term which specificies the curvature of the intensity functions, that is the square of the second derivates. The penalized log-likelihood (pl) is defined as:

$$pl = l - \kappa_{01} \int \alpha_{01}''^{\,2}(u|Z_{01})du - \kappa_{02} \int \alpha_{02}''^{\,2}(u|Z_{02})du - \kappa_{12} \int \alpha_{12}''^{\,2}(u|Z_{12})du \qquad (4)$$

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. The smoothing parameters can be chosen either arbitrarily or by maximizing a cross-validation score. The leave-one-out cross-validation involves using a single observation as the validation data and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as validation data. To avoid maximizing the likelihood as many times as there are observations in the data set (and for each different values of κ_{01}, κ_{02}, κ_{12}), we use an approximate leave-one-out cross-validation score proposed by O'Sullivan (1988) for survival models and extended by Commenges *et al.* (2007) to multi-state models. It requires maximizing the likelihood one time only by tested values of the smoothing parameters. To find the values of κ_{01}, κ_{02}, κ_{12} which maximize this score we use a grid search method.

For given smoothing parameters, maximization of (4) defines the maximum penalized likelihood estimators of the baseline transition intensities $\hat\alpha_{0,01}$, $\hat\alpha_{0,02}$ and $\hat\alpha_{0,12}$ which are approximated using a basis of M-splines. The parameters being maximized are the regression coefficients and the coefficients of the linear combination of M-splines.

The splines approximation

A M-spline (Ramsay 1988) is a non negative spline. A family of M-spline functions of order k, $M_1, \ldots, M_n$ is defined by a set of m knots $t = (t_1 \leq t_2 \leq \ldots \leq t_m)$ where $n = m + k - 2$. We use cubic M-splines, i.e. M-splines of order $k = 4$.

We denote by $t_{01} = (t_{01,1}, \ldots, t_{01,m_{01}})$ the sequence of m_{01} knots used to define the cubic M-splines approximation of $\hat\alpha_{0,01}$, and by $t_{02} = (t_{02,1}, \ldots, t_{02,m_{02}})$ and $t_{12} = (t_{12,1}, \ldots, t_{12,m_{12}})$ similar sequences for $\hat\alpha_{0,02}$ and $\hat\alpha_{0,12}$, respectively. We denote by $M_{hl}^T = M_{hl,1}, \ldots, M_{hl,n_{hl}}$ the matching families of n_{hl} cubic M-splines, with $n_{hl} = m_{hl} + 2$.

For $hl \in \{01, 02, 12\}$, the estimator $\hat\alpha_{hl}$ is approximated using the linear combination:

$$\tilde\alpha_{0,hl}(x) = \sum_{i=1}^{n_{hl}} a_{hl,i} M_{hl,i}(x)$$

where $a_{hl,i}$ are coefficients to estimate.

Non-negativity of $\tilde\alpha_{0,hl}$ is obtained by constraining the coefficients $a_{hl,i}$ to be positive. In practice, we estimate parameters $\theta_{hl,i}$ such that $a_{hl,i} = \theta_{hl,i}^2$.

The n_{hl} M-splines can be integrated to produce a family of monotone splines called I-splines. With each M-spline $M_{hl,i}$ we associate an I-spline $I_{hl,i}$:

$$I_{hl,i}(x) = \int_{t_{hl,1}}^{x} M_{hl,i}(u)du$$

Given the coefficients $a_{hl,i}$, we can approximate the estimators of the cumulative baseline transition intensities $\hat A_{hl}$ by a linear combination of I-splines:

$$\tilde A_{0,hl}(x) = \sum_{i=1}^{n_{hl}} a_{hl,i} I_{hl,i}(x).$$

Because M-splines are non-negative, the positivity constraint on $a_{hl,i}$ ensures that $\tilde{A}_{0,hl}$ is monotone increasing.

Confidence intervals of the regression parameters are obtained using estimated standard errors estimated by inverting the Hessian matrix of the log-likelihood. Confidence intervals for the transition intensities $\alpha_{hl}(t)$ are obtained using a Bayesian approach proposed by O'Sullivan (1988) for survival analysis where the standard errors are estimated by $M_{hl}(t)^T H^{-1} M_{hl}(t)$ with H the Hessian matrix of the penalized log-likelihood. (PIERRE: is it correct ?)

4. Predictions

Most often in illness-death models, the functions of interest are the transition intensities. However, other quantities which can be expressed in terms of the transition intensities (Touraine *et al.* 2013) may provide additional information and have a more natural interpretation.

For example, given a set of covariates $Z_{01,i}, Z_{02,i}, Z_{12,i}$ for a subject i who is diseased at time s, one could be interested in probability to be still alive at some time $t > s$, or in life expectancy; given a set of covariates $Z_{01,j}, Z_{02,j}, Z_{12,j}$ for a subject j who is diseased-free at time s, one could be interested in lifetime risk of disease or in healthy life expectancy (expected remaining sojourn time in state 0). Since these quantities can be written in terms of the transition intensities, **SmoothHazard** provides estimates of them using estimates of the transition intensities. Confidence intervals of these quantities are calculated using the simulation-based method immediately following.

4.1. Confidence regions

A simulation based approach (Mandel 2013) is used to calculate confidence intervals for the transition intensities $\alpha_{hl}(t)$ in the parametric approach and for the quantities of interest (transition probabilities, cumulative probabilities and life expectancies) in both parametric and semi-parametric approaches.

We assume the asymptotic normality for the estimator $\hat{\theta}$ (either for the regression parameters and the distribution parameters in the parametric approach, or for the regression parameters and the spline parameters in the semi-parametric approach). We denote by $\hat{V}_{\hat{\theta}}$ the estimated covariance matrix of $\hat{\theta}$. We consider a multivariate normal distribution with the parameters estimates as expectation and $\hat{V}_{\hat{\theta}}$ as covariance matrix. We generate n vectors ($n = 2000$ in practice) from this distribution: $\theta^{(1)}, \dots, \theta^{(n)}$. Based on them, we can calculate n values for the transition intensities: $\alpha_{hl}^{(1)}(t)$,...,$\alpha_{hl}^{(n)}(t)$, and therefore n values for any quantity of interest written in terms of the transition intensities. The n values reflecting the sample variation (Aalen *et al.* 1997), we order them and the 2.5^{th} and the 97.5^{th} empirical percentiles are then used as lower and upper confidence bounds for 95% confidence intervals. This procedure can be repeated for any t, so we can obtain pointwise confidence bands for $\alpha_{hl}(.)$.

5. Using SmoothHazard

5.1. Data and main arguments

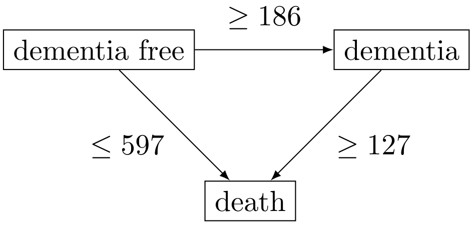


Figure 2: The exact number of transitions in the illness-death model with interval-censored time to disease is unknown.

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.

```
library(SmoothHazard)
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 between the diagnostic visit and the previous one and demented subjects are at risk of death. Thus, subjects who died without being diagnosed as demented at their last visit may have become demented between last visit and death.

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 `127` should have been dead with dementia and less than `597` should have been dead without dementia (see Figure 5.1).

Age is chosen as the basic time scale and subjects are dementia-free (and alive) at entry into study. 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 if an event occurred, 0 otherwise) for dementia and death, respectively. The variable `t0` contains ages of subjects at entry into study. 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 on vital status. There are two binary covariates: `certif` for primary school diploma (`762` with diploma and `238` without diploma) and `gender` (`578` women and `422` men).

The function `idm` computes estimates for the three transition intensities $\alpha_{01}(.)$, $\alpha_{02}(.)$, $\alpha_{12}(.)$ which are age-specific incidence rate of dementia, age-specific mortality rate of dementia-free subjects and age-specific mortality rate of demented subjects, respectively. Proportional transition intensities regression models allow for covariates on each transition. Covariates are specified independently for the regression models of the three transition intensities by the right hand side of the respective formula `formula01`, `formula02` and `formula12`.

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

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

The main function `idm` computes estimates for the three baseline transition intensities and for the regression parameters of an illness-death model. The `intensities` argument by specifying the form of the transition intensities allows to select either a parametric or a semi-parametric estimation method :

- With the default value `"Weib"`, a Weibull distribution is assumed for the baseline transition intensities and the parameters are estimated by maximizing the log-likelihood;
- With the `"Splines"` value, the estimation is conducted by maximizing a penalized log-likelihood where the transition intensities estimators are approximated by linear combinations of M-splines.

We stop the iterations of the maximization algorithm 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.

We now illustrate how to fit the illness-death model to the `Paq1000` data set, based on interval-censored dementia times and exact death times.

In the following call, a Weibull parametrization is used for the three baseline transition intensities and we include two covariates on the transition to dementia, one covariate on the transition from no dementia to death and no covariates on the transition from dementia to death. Note that in case of missing `formula12` argument the covariates on the $1 \rightarrow 2$ transition are the same as the ones specified in the `formula02` argument.

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

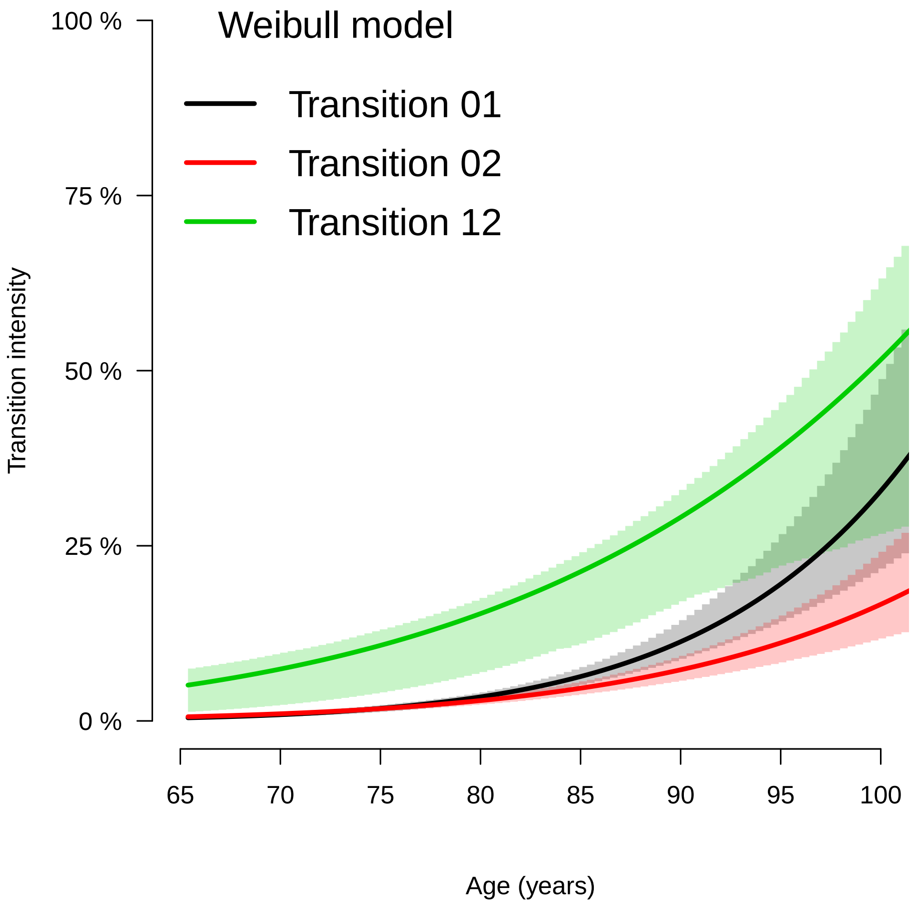
```
 Erreur : symbole inattendu(e) in:
"pdf(file="fig1.pdf")
library"
null device
          1
[1] 186
[1] 724
[1] 597
[1] 186
[1] 127
[1] 597
[1] 762
[1] 238
[1] 578
[1] 422
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = t0) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = t0) ~ gender, formula12 = ~1, data = Paq1000)

Illness-death model: Results of Weibull regression for the intensity functions.

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

                coef SE.coef     HR          CI      Wald p.value
certif_01_01 -0.4117  0.1827 0.6625 [0.46;0.95]  5.077106 0.02424
gender_01_01 -0.2621  0.1561 0.7694 [0.57;1.04]  2.818364 0.09319
gender_02_02  0.6712  0.1143 1.9565 [1.56;2.45] 34.449583 < 1e-04

               Without cov  With cov
Log likelihood   -3075.308 -3053.648

Parameters of the Weibull distribution: 'S(t) = exp(-(b*t)^a)'
      alpha01    alpha02    alpha12
a 11.12344625 8.82268159 6.44006486
b  0.01102198 0.01074539 0.01381268
```

```
----
Model converged.
number of iterations:  6
convergence criteria: parameters= 7.3e-10
                    : likelihood= 2.3e-08
                    : second derivatives= 2.8e-12
```

The hazard ratios HR (e^{coef}) have the usual interpretation, as in a parametric Cox regression model.

The three baseline transition intensity functions can be displayed as functions of time, functions of age in our illustrative example (Figure 3).

```
par(mgp=c(4,1,0),mar=c(5,5,5,5))
plot(fit.weib,conf.int=TRUE,lwd=3,citype=”shadow”,xlim=c(65,100), axis2.las=2,axis1.at=seq
    (65,100,5),xlab=”Age (years)”)
```



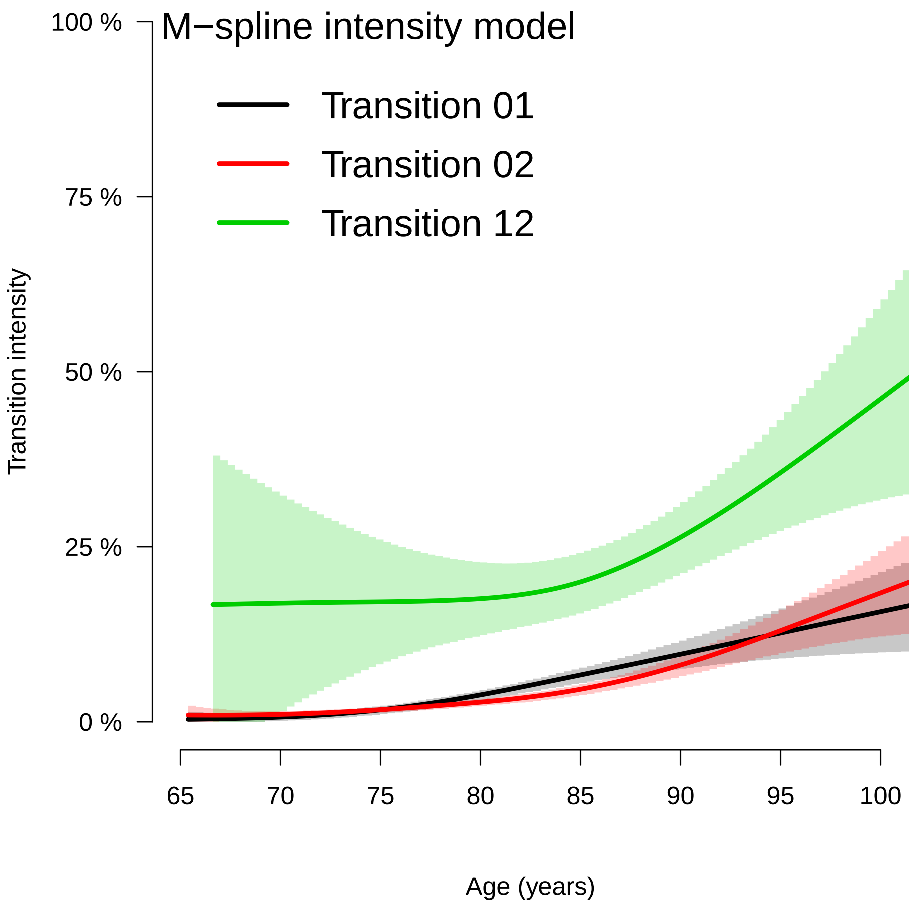
The other estimation option in the function `idm` permits to relax the strict parametric assumptions of the Weibull regression models. With the option `intensities="Splines"`, linear combinations of M-splines are used to approximate the three baseline transition intensities. Although this option implies a considerable amount of extra computations (see Section 3.2), the call and the printed output are very similar to the Weibull model:

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

```
null device
          1
Call:
idm(formula01 = Hist(time = list(l, r), event = dementia, entry = t0) ~
    certif + gender, formula02 = Hist(time = t, event = death,
    entry = t0) ~ gender, formula12 = ~1, data = Paq1000, intensities = "Splines")

Illness-death regression model using M-spline approximations
 of the baseline transition intensities.

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

Smoothing parameters:
      transition01 transition02 transition12
knots        7e+00        7e+00            7
kappa        8e+05        2e+05        50000

                coef SE.coef     HR           CI      Wald p.value
certif_01_01 -0.3759  0.1853 0.6867 [0.48;0.99]  4.115764 0.04249
gender_01_01 -0.2299  0.1580 0.7946 [0.58;1.08]  2.116337 0.14573
gender_02_02  0.6536  0.1120 1.9225 [1.54;2.39] 34.081159 < 1e-04

                         Without cov  With cov
Penalized log likelihood   -3072.387 -3051.939

----
Model converged.
number of iterations:  9
convergence criteria: parameters= 2.9e-08
                    : likelihood= 6.7e-07
                    : second derivatives= 1.7e-10
```

Again, the estimated baseline transition intensities can conveniently be visualized in a joint graph (Figure 4).

```
par(mgp=c(4,1,0),mar=c(5,5,5,5))
plot(fit.splines,conf.int=TRUE,lwd=3,citype=”shadow”,xlim=c(65,100), axis2.las=2,axis1.at=seq(65,100,5),xlab=”Age (years)”)
```



Semi-parametric estimation method: specific options

Some optional arguments are specific to the semi-parametric approach (when using the option `intensities="Splines"`:

- `n.knots` contains a vector (by default `c(7,7,7)`) specifying the number of knots on the 0 → 1, 0 → 2 and 1 → 2 transitions, respectively;

- `knots` contains the choice of the knots placement (equidistant by default or quantile-based placement) or a list of sequences of knots for transitions 0 → 1, 0 → 2 and 1 → 2, respectively, to be specified by the user;

- `CV` (FALSE by default) is set to TRUE for using approximate leave-one-out cross-validation score to choose the smoothing parameters κ_{01}, κ_{02}, κ_{12};
- `kappa` contains the smoothing parameters if `CV=FALSE` (arbitrary choice of the smoothing parameters κ_{01}, κ_{02}, κ_{12}); the initial smoothing parameters for the grid search method which maximize the approximate leave-one-out cross-validation score if `CV=TRUE`.

By default the function `idm` selects equidistant sequences of 7 knots between the minimal and maximal event times (`t0`, `l` and `r` for `Paq1000`). There must be several data points between each pair of different knots and there must be a knot before or at the first time from which there are subjects at risk and after or at the last time of transition. Five is the minimal number of knots that can be chosen for a transition. Indeed, the penalized likelihood approach has no interest with very few knots (PIERRE: Can you develop/give a better explanation, or do I delete ?). Consequently, the semi-parametric approach requires much more information than the parametric one to achieve convergence. The number of parameters to be estimated is larger, and enough observation times on each transition are required to fit the splines. In particular, in data sets where few 1 → 2 transitions times are observed, we this approach is not recommended. Increasing the number of knots does not deteriorate the estimates of the transition intensities: 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 (from our own experience rarely more than 12 knots).

The default values for the smoothing parameters are suitable for the `Paq1000` data set. However, these values can be expected to be very different depending on time scale, number of subjects and number of knots. The cross-validation option can be used to find appropriate smoothing parameters. However, the running time with cross-validation is very long and an empirical technique can be preferred. It consists in repeating the `idm` running trying different smoothing parameters. After each estimation, the transition intensities are plotted. This can be done with the `plot` function. If the curves seem too smooth it may be useful to reduce the associated smoothing parameter. Similarly, if the curves are to wiggly, the associated smoothing parameter may be increased.

5.3. Making predictions

The function `idm` returns an "idmWeib" or "idmSplines" class object depending on the parametrization of the transition intensities (Weibull or splines). A object as returned by the `idm` function can be used in argument of the `predict` function in order to obtain transition probabilities, cumulative probabilities of event and life expectancies with confidence intervals. For example, the following call give predictions regarding a 70 years-old female subject who do not have primary school diploma, over a 10 years horizon:

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

```
null device
          1
$p00
[1] 0.6351952 0.5907158 0.6776652

$p01
[1] 0.04764996 0.03326921 0.07342708

$p11
[1] 0.3337669 0.2650398 0.6707413

$p12
[1] 0.6662331 0.3292587 0.7349602

$p02_0
[1] 0.2872955 0.2409324 0.3322761

$p02_1
[1] 0.02985929 0.01215399 0.04579234

$p02
[1] 0.3171548 0.2743960 0.3577830

$F01
[1] 0.07750925 0.05104802 0.11713811

$F0.
[1] 0.3648048 0.3223348 0.4092842
```

The covariates values must be specified in the `Z01`, `Z02` and `Z12` arguments in the same order as they were entered in the preceding `idm` call.

The ouput attributes are:

- for a dementia-free 70 years-old subject:
  - the probability of being still alive and dementia-free 10 years later $p_{00}(70, 80)$,
  - the probability of being still alive but demented 10 years later $p_{01}(70, 80)$,
  - the probability of dying in the next 10 years $p_{02}(70, 80)$ having been demented before ($p^1_{02}(70, 80)$) or not ($p^0_{02}(70, 80)$),
  - the absolute risk of dementia in the 10 years (10 years later, the subject may have die or not) $F_{01}(s, t)$,
  - the absolute risk of exit from state 0 in the 10 years $F_{0\bullet}(s, t)$ (due to either dementia or death);
- for a demented 70 years-old subject: the probability of dying in the next 10 years $p_{12}(s, t)$ or not $p_{11}(s, t)$.

The following calls give life expectancies regarding a 70 years-old female subject who do not have primary school diploma based on the transition intensities estimates from respectively the parametric approach and the semi-parametric approach:

```
LE.weib <- lifexpect(fit.weib,s=80,Z01=c(1,0),Z02=0)
LE.weib
```

```
$life.in.0.expectancy
[1] 8.868163 7.893798 9.782593

$life.expectancy.nondis
[1] 10.445056  9.790562 11.605361

$life.expectancy.dis
[1] 4.890873 4.402014 7.866383
```

```
LE.splines <- lifexpect(fit.splines,s=80,Z01=c(1,0),Z02=0,CI=FALSE)
LE.splines
```

```
$life.in.0.expectancy
[1] 8.818227

$life.expectancy.nondis
[1] 10.4169

$life.expectancy.dis
[1] 4.90946
```

The confidence intervals calculation may take time, especially using the splines estimates of the transition intensities. To suppress this calculation the `CI` argument must be set to `FALSE` (see above). The number of the simulations for calculating confidence intervals can also be modified using the `nsim` argument (by default 2000 for the `predict` function and 1000 for the `lifexpect` function).

The output attributes of the `lifexpect` function are:

- for a dementia-free 80 years-old subject:
  - the life expectancy in state 0 (healthy life expectancy),
  - the life expectancy;
- for a demented 80 years-old subject: the life expectancy.

Warnings regarding predictions

Predictions using the splines estimates of the transition intensities are not possible if involving times prior to the first knot or times beyond the last knot. Moreover, the life expectancies are calculated using integration until infinity using the Weibull estimates and until the last knot using the splines estimates. Consequently, to calculate life expectancies using the splines estimates, we implicitly assume that the last knot time is the maximal time of death. The above life expectancies calculating from the Weibull estimates or the splines estimates of the transition intensities are very close because the follow-up period of the `Paq1000` data set is long. However, in other data sets this assumption may not hold anymore. Finally, to avoid numerical problem in the predictions calculations, the first and last knots must be the same or very close on each transition.

References

Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON (1997). "A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales." *Statistics in Medicine*, **16**(19), 2191–2210.

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.

Jackson C (2011). "Multi-State Models for Panel Data: The msm Package for R." *Journal of Statistical Software*, **38**(8), 1–28.

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.

Leffondré K, Touraine C, Helmer C, Joly P (2013). "Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model?" *International journal of epidemiology*.

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.

Levenberg K (1944). "A method for the solution of certain problems in least squares." *Quarterly of applied mathematics*, **2**, 164–168.

Mandel M (2013). "Simulation Based Confidence Intervals for Functions with Complicated Derivatives." *The American Statistician*, **67**.

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/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd