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,

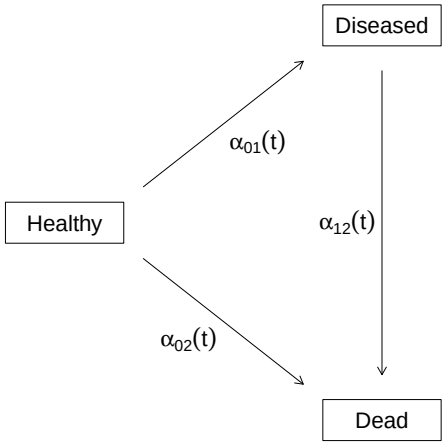


Figure 1: The irreversible illness-death model has three transition intensities.

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 [REF: Joly et al.].

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

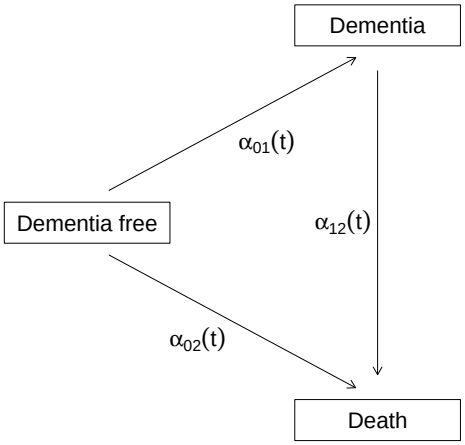
Section 2 presents the models and the likelihood. Section 3 presents the estimation methods. Section 4 provides some examples illustrating **SmoothHazard** functions.

1.1. Data

Paquid is a large cohort study on mental and physical aging. The population consists of subjects aged 65 years and older living in Southwestern France. 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). The event of interest is the incidence of 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. There are two covariates in this subset: sex (`578` women and `422` men) and primary school diploma (`762` with diploma and `762` without diploma). Age is chosen as the basic time scale. Consequently, we need to deal with left-truncated event times.

```
1 library(prodlim)
2 plotIllnessDeathModel(stateLabels=c("Dementia free","Dementia","Death"),arrowLabelSymbol="
      alpha")
```




```
head(Paq1000)
```

```
  dementia death   entry        L        R     time 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
```

1.2. Questions

TODO

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

The function idm computes the maximum likelihood estimate for the three transition intensities:

In the situation where both transition times are not interval censored the regression coefficients can be estimated by the partial likelihood method **?** 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 transition intensities simultaneously and that we need a model for the baseline hazard functions. This can be seen by inspecting the likelihood function. Denote the conditional event-free survival function by

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

where the conditional cumulative hazard

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

.

Subject i, δ_{1i} indicator for event "ill", δ_{2i} indicator for event "death", cumulative intensity function. We can have $L_i = R_i$: no interval censoring, $L_i = R_i = T_i$ if we know that the subject is not becoming ill before dying, $R_i = T_i$ if we do not know that the subject is becoming ill before dying

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

2.1. The Weibull parametrization

TODO: Describe the implemented parametrization of the three baseline functions.

```
fit.weib <- idm(formula02=Hist(time,event=death,entry=entry)~certif+gender,
                formula01=Hist(time=list(L,R),event=dementia)~certif+gender,
                data=Paq1000,eps=c(5,5,3),maxiter=200,hazard="Weib")
print(fit.weib)
```

```
Call:
idm(formula01 = Hist(time = list(L, R), event = dementia) ~ certif +
    gender, formula02 = Hist(time, event = death, entry = entry) ~
    certif + gender, data = Paq1000, maxiter = 200, eps = c(5,
    5, 3), hazard = "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]  6.6399364 0.009972
gender_01 -0.1221  0.1599 0.8851 [0.65;1.21]  0.5834324 0.444970
certif_02  0.1268  0.1264 1.1352 [0.89;1.45]  1.0066517 0.315706
gender_02  0.5363  0.1200 1.7096 [1.35;2.16] 19.9873828 < 0.0001
certif_12 -0.2079  0.2323 0.8123 [0.52;1.28]  0.8014211 0.370669
gender_12  0.5792  0.1865 1.7846 [1.24;2.57]  9.6469569 0.001897

               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 11.18802187 8.62750163 7.50200265
b  0.01099806 0.01078284 0.01294115

----
Model converged.
number of iterations:  8
convergence criteria: parameters= 0.00000012
                    : likelihood= 0.0000007
                    : second derivatives= 0.00000000047
```

2.2. The penalized likelihood

Intensity functions are expected to be smooth.

To introduce such a priori knowledge, we penalize the likelihood by a term which has large values for rough functions.

The roughness penalty function chosen for the three-state model is the sum of the square norms of the second derivatives of the intensities.

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

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

where l is the full log-likelihood (which is a function of $\alpha_{01}(.)$, $\alpha_{12}(.)$ and $\alpha_{02}(.)$) and κ_{01}, κ_{12} and κ_{02} are three positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions.

Maximization of (2) defines the maximum penalized likelihood estimators (MPLE) $\hat{\alpha}_{01}(.)$, $\hat{\alpha}_{12}(.)$ and $\hat{\alpha}_{02}(.)$.

Approximation via splines:

The MPLE of (2) cannot be calculated explicitly. However, it can be approximated using splines.

Splines are piecewise polynomial functions which are combined linearly to approximate a function on an interval.

We use cubic M-splines and I-splines, which are variants of B-splines.

The estimator $\hat{A}(.)$ for a given transition is approximated by a linear combination of m I-splines:

\$ {˜ A}(.) = $\tilde{\theta}$I (.)\$, where $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_m)$ and $I(.) = (I_1(.), ..., I_m(.))^T$. By differentiation we obtain: $\tilde{\alpha}(.) = \tilde{\theta} M(.)$, where $M(.) = (M_1(.), ..., M_m(.))^T$. We use a distinct base of splines for each intensity function, possibly with a different number of splines in each basis. The monotonicity constraint for \$ {˜ A}(.)\$ is fulfilled by constraining the coefficients $\tilde{\theta}$ to be positive.

The approximation $\tilde{\alpha}$ of $\hat{\alpha}$ is the function belonging to the space generated by the basis of splines, which maximizes $pl(\alpha_{01}, \alpha_{12}, \alpha_{02})$.

We briefly present the M-splines and I-splines used here and give some computational aspects of this approach.

For more details see Ramsay (1988).

A M-spline of order k is defined as:

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

with

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

where $t_1, ..., t_m$ is a sequence of increasing knots.

Each $M_j(x|k)$ is zero outside of the interval $[t_j, t_{j+k}]$, hence is non-zero over k intervals and over each interval there are k non-zero M-splines. For our approximation we use splines of order 4 (cubic splines).

To each M-spline we associate a I-spline:

$$I_j(x|k) = \int_0^x M_j(u|k)du.$$

Each M_j is piecewise polynomial of degree $k-1$ and each associated I_j is piecewise polynomial of degree k defined as (if $t_j \leq x < t_{j+1}$):

$$I_h(x|k) = \begin{cases} 0 & \text{if } h > j, \\ \sum_{l=h}^{j}(t_{l+k+1} - t_l)\frac{M_l(x|k+1)}{k+1} & \text{if } j-k+1 \leq h \leq j, \\ 1, & \text{if } h < j-k+1. \end{cases}$$

These splines are convenient to manipulate; among other things a linear combination of splines is easy to differentiate or integrate.

Note that M-splines are nonnegative and I-splines are monotonically increasing; it results that 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 $\hat{A}(.)$ can be approximated by a linear combination of m I-splines \$ {˜ A}(.) = $\sum_{\text{j=1}}$ {̂m} g(˜ θ _j)I _j (.)\$, where $g(\tilde{\theta}_j) \geq 0 \ \ \forall j$ (for example $g(\tilde{\theta}_j) = e^{\tilde{\theta}_j}$ or $g(\tilde{\theta}_j) = \tilde{\theta}_j^2$); in practice we use $g(\tilde{\theta}_j) = \tilde{\theta}_j^2$ to avoid convergence problems when $g(\tilde{\theta}_j)$ should be zero. For the transition intensity we have: \${˜ α}(.) = $\sum$ _{j=1} m̂ g(˜ θ _j) M _j(.) .\$ So with the same vector of coefficients $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_m)^T$ we get the cumulative hazard function with I-splines and the hazard function with M-splines. In fact the set of functions generated by the basis of splines with positive coefficients is included in the set of positive functions generated by the basis of splines. However our numerical experience shows that this set is rich enough to provide a good approximation of the hazard function.

{The knots :}

A spline function is completely defined by a sequence of knots and the coefficients of the splines.

In the program, a knot is set on the first and last data points and the other knots are put equidistantly between them by default.

Another way to have an automatic choice for the location of the knots is to locate the knots at every p data points as described in O'Sullivan (1988). Otherwise the user can choose their location freely but by verifying that there are several time in the data set between every knots.

Theoretically, the more knots, the better the approximation.

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} 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 parameters; 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). It is possible to have different number of knots for each transition intensity.

In any case there must be a knot before or at the first data point and after or at the last data point.

%%%%%%%%%%%%%% Penalized

The vectors of spline coefficients $\tilde{\theta}_{01}$, $\tilde{\theta}_{12}$ and $\tilde{\theta}_{02}$ for fixed κ_{01}, κ_{12} and κ_{02} are obtained simultaneously by maximizing the log-likelihood using a Marquardt's algorithm (1963)

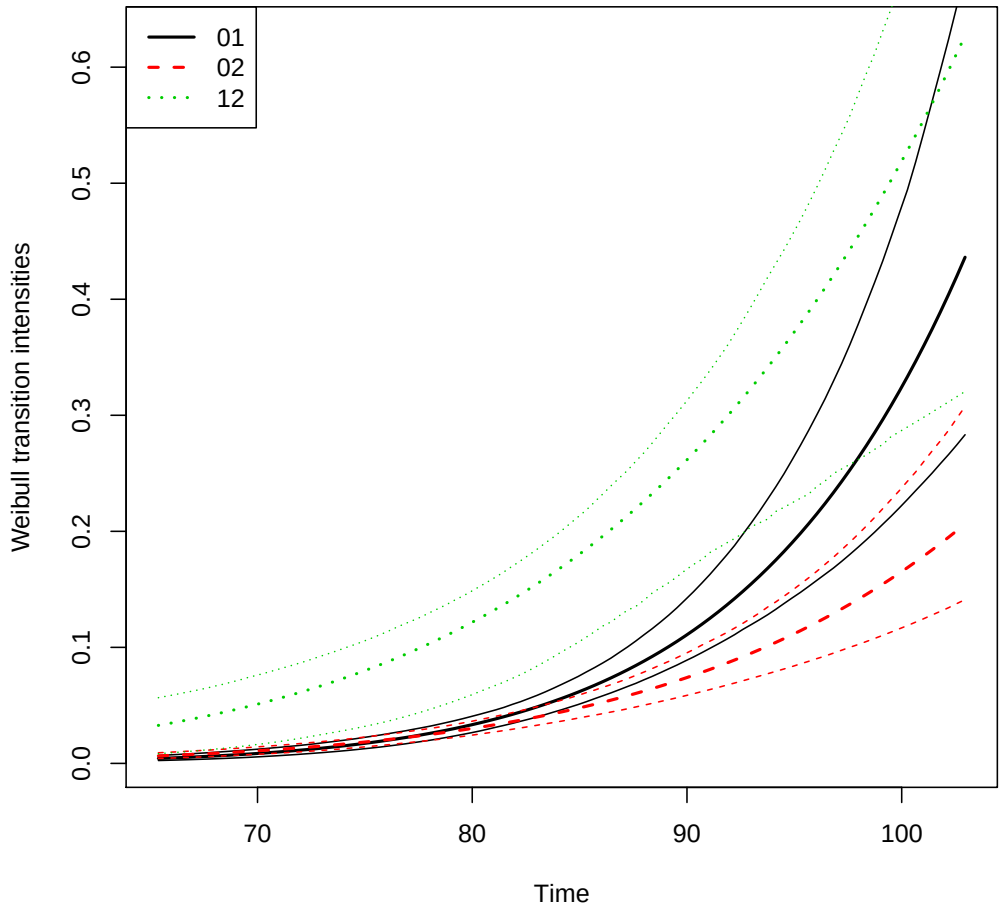
When the three vectors $\tilde{\theta}_{01}$, $\tilde{\theta}_{12}$ and $\tilde{\theta}_{02}$ are obtained, with the knots sequence , all the functions of interest can be computed, as in a parametric method.

{Algorithm:}

The vectors of parameter for the baseline transition intensities (either spline coefficients or weibull parameters) and the vector of regression parameter $\hat{\beta}_{01}$, $\hat{\beta}_{12}$ and $\hat{\beta}_{02}$ are obtained simultaneously by maximizing the log-likelihood using a combination of a Marquardt's algorithm (1963) and a steepest descent algorithm. Marquardt's algorithm is a robust Newton-like algorithm. The Marquardt's algorithm step involves a line search with a step reduction if the new point is not better. 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 of the log-likelihood. Few iterations are needed if the initial value is judiciously chosen because the Marquardt's algorithm iteration is used. In other cases the steepest descent iteration is often used because the Hessian may be singular and the convergence is slower.

We stop the iterations when the difference between two consecutive log-likelihoods is small, the coefficients are stable and the gradient is small enough. The variance of parameter estimates

are estimated using the inverse of the matrix of the second derivatives at convergence.



3. Predicting parameters of life

References

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.

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.

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