A gap time model based on restricted cubic splines with a zero-recurrence proportion

Ivo Sousa-Ferreira[1], Cristina Rocha[1], Ana Maria Abreu[2]

[1]DEIO & CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal

[2]DM & CIMA, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal

ivo.ferreira@staff.uma.pt; cmrocha@fc.ul.pt; abreu@staff.uma.pt

Introduction

Recurrent events data arise frequently in medical studies where each subject may experience a particular event repeatedly over time. Recently, considerable attention has been devoted to modelling the gap times. Here, we consider the classic assumption that the number of recurrent events up to a given time follows a NHPP, for which the gap times are generally not independent. In this sense, we follow the approach of Zhao and Zhou [3], wherein the recurrence process is derived from a NHPP. However, we assume a completely parametric baseline rate function in which the covariates have a multiplicative effect.

The main challenge here is to select the most appropriate baseline form, which sometimes is not flexible enough to capture how the rate evolves over time. Motivated by Royston and Parmar [2], we propose to use restricted cubic splines (RCS) to overcome this shortcoming.

Restricted cubic splines

A RCS function is a collection of piecewise cubic polynomials joined at a pre-defined number of internal knots, that it is also constrained to be linear beyond the boundary knots to ensure a sensible form (see Figure 1).

For pre-defined m distinct internal knots $r_1 < \ldots < r_m$, with $r_{\min} < r_1$ and $r_{\max} > r_m$ boundary knots, the RCS function of a given observed variable x may be written as

$$s(x;\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \sum_{l=1}^{m} \gamma_{l+1} v_l(x), \qquad (1)$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{m+1})'$ is the parameters vector and $v_l(x)$ is known as the lth basis function.

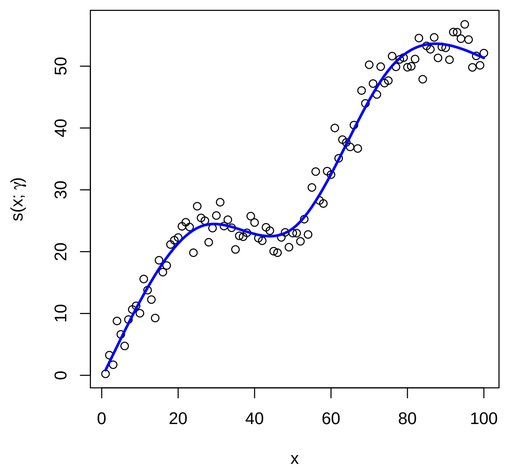

Figure 1: Estimated RCS function for pseudo-random data generated from a 4th degree polynomial plus error from a normal distribution with mean 0 and standard deviation 2.5.

Notation

- There are n independent subjects in study and each one experiences K_i ($i = 1, \ldots, n$) recurrences of an event;
- T_{ik} is the time since the beginning of the study until the occurrence of the kth event ($k = 1, \ldots, K_i$);
- $Y_{ik} = T_{ik} - T_{i,k-1}$, with $0 \equiv T_{i0} < T_{i1} < \ldots < T_{iK_i}$, is the gap time between two consecutive events of the ith subject;
- $\boldsymbol{z}_{ik} = (z_{ik1}, \ldots, z_{ikp})'$ is a vector of covariates for the ith subject with respect to the kth event and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients (latency part of the model);
- π is the probability of being a recurrent subject (susceptible) and $1 - \pi$ is the probability of being a zero-recurrence subject (non-susceptible);
- It is natural to assume that π can be written in terms of the covariates via a logistic function. In our case, it follows that $\pi_i = 1/[1 + \exp(-\boldsymbol{\alpha}' z_{i1})]$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ is a vector of regression coefficients (incidence part of the model).

Model formulation

Based on Zhao and Zhou [3], the recurrence process is assumed to be a NHPP with independent increments. Then, we consider a multiplicative model in which the marginal rate function is given by

$$h(y|t_{i,k-1}, \boldsymbol{z}_{ik}) = h_0(y + t_{i,k-1}) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ik}), \qquad (2)$$

where $h_0(\cdot) > 0$ is a baseline rate function.

Following the approach of Royston and Parmar [2], we propose to model the log-cumulative baseline rate function as a RCS function of log time, which provides analytically tractable expressions. From (2), the cumulative rate function is

$$\begin{aligned} H(y|t_{i,k-1}, \boldsymbol{z}_{ik}) &= \left[\exp\left\{ \log H_0(y + t_{i,k-1}) \right\} - \exp\left\{ \log H_0(t_{i,k-1}) \right\} \right] \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ik}) \\ &= \left[\exp\left\{ s\big(\log(y + t_{i,k-1}); \boldsymbol{\gamma} \big) \right\} - \exp\left\{ s\big(\log t_{i,k-1}; \boldsymbol{\gamma} \big) \right\} \right] \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ik}), \end{aligned}$$

where $H_0(\cdot) > 0$ is a cumulative baseline rate function and $s(\log t; \boldsymbol{\gamma})$ is the RCS function (1) of log time.

In some scenarios, it might exist a proportion of the population under study that becomes recurrence free. Therefore, we consider that two cases can occur:

- if $K_i > 1$, subject i experiences at least one recurrence, so he is a recurrent subject;
- if $K_i = 1$, subject i may either be a recurrent subject with probability π_i or a zero-recurrence subject with probability $1 - \pi_i$.

The inferential procedure is based on the maximum likelihood (ML) method, assuming a non-informative right-censoring mechanism. For each subject i, we define $\delta_i = I(K_i > 1)$ and $K_i^* = \max(K_i - 1, 1)$. The likelihood function is expressed as

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ \pi_i \prod_{k=1}^{K_i^*} f(y|t_{i,k-1}, \boldsymbol{z}_{ik}) \right\}^{\delta_i} \left\{ 1 - \pi_i + \pi_i P(Y_{i1} > y | T_{i0} = 0) \right\}^{1-\delta_i},$$

where $f(y|t_{i,k-1}, \boldsymbol{z}_{ik})$ is the probability density function and $P(Y_{i1} > y|T_{i0} = 0)$ is the (proper) survival function of the first gap time. The computational implementation was developed in R software [1], version 4.1.0, where the ML estimates were obtained using the Broyden–Fletcher–Goldfarb–Shanno maximization procedure.

An application to re-hospitalization data

The re-hospitalizations data represents the gap times (in days) of successive readmissions of 403 patients diagnosed with colorectal cancer after receiving surgery to remove their tumours. The maximum follow-up time was 2176 days (about 6 years). A total of 861 readmissions were recorded, ranging from 1 to 22, with mean 2.3 and median 1.0. About 49.4% of the patients had no recurrence at all (see Table 1). The data are available in the `R` [1] library `frailtypack`.

Table 1: Information on the first 4 recurrences.

Number of subjects	Recurrence number			
	1	2	3	4
At risk	403	204	99	54
Who experienced	204	99	54	33
% of censoring	49.4	51.5	45.5	38.9

Here, 4 covariates were included in the model: chemotherapy; gender; Dukes' stage; and Charlson comorbidity index. In preliminary modelling, without covariates and zero-recurrence proportion, we use the Akaike (AIC) and Bayesian (BIC) information criteria to guide the choice of number of d.f. required to capture the baseline rate function. Thus, models with between 1 to 4 d.f. were fitted. The AIC values (6883.1, 6871.3, 6872.4 and 6872.9) and BIC values (6892.6, 6885.5, 6891.4 and 6896.7) indicate that 2 d.f. (1 internal knot) is the most adequate choice. Then, the proposed flexible model that accounts for zero-recurrence subjects was applied and the results are summarized in Table 2.

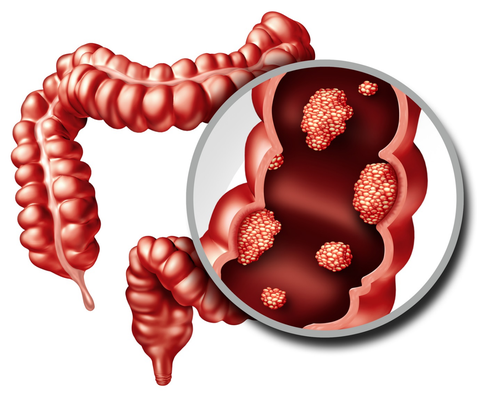

Figure 2: Illustration of colorectal cancer. Image adapted from: `https://tinyurl.com/colonrcancer`.

Table 2: Parameters estimates of the flexible marginal rate model with 1 internal knot and a zero-recurrence proportion.

	Parameter	Estimate	$\widehat{SE}$	p-value
Rate component (latency)	γ_0	-5.868	0.642	–
	γ_1	1.019	0.170	–
	γ_2	0.001	0.005	–
	Chemo	-0.048	0.115	0.677
	Gender	-0.444	0.111	<0.001
	Dukes' stage			
	C	0.198	0.126	0.116
	D	0.641	0.147	<0.001
	Charlson index			
	1–2	0.357	0.208	0.087
	3	0.528	0.117	<0.001
Logistic component (incidence)	Intercept	0.180	0.268	0.500
	Chemo	-0.341	0.266	0.200
	Gender	-0.163	0.248	0.511
	Dukes' stage			
	C	0.329	0.261	0.207
	D	1.979	0.593	<0.001

In our model, the reference group consists of male patients who did not receive chemotherapy, with Dukes' stage A−B and Charlson index 0. For this group the zero-recurrence proportion is $1-(1/\{1+\exp(-0.180)\}) = 0.455$. The chemotherapy coefficient estimates are negative in both rate and logistic components with a non-significant effect on the time to readmission. This suggest that, although chemotherapy diminishes the rate of readmission, its effect on the readmission process is negligible. In the rate component, recurrent females have a significantly lower risk of readmission compared with recurrent males. The other two important risk factors in this component are the Dukes' stage D and the Charlson index ≥ 3. In relation to the logistic component, only Dukes' stage D has a significantly increasing effect, which means that patients in this stage have lower chances of being recurrence free.

The Cox-Snell residuals were used to informally assess the overall goodness of fit of the model. From the left plot of Figure 3 it is confirmed that the model provides a good fit of the data, since the residuals behave as a straight line through the origin with slope 1. The model-based estimate of the marginal rate function is depicted in the right plot of Figure 3, which exhibits a right-skewed unimodal shape. This means that the re-hospitalization rate increases during the first 30 days after the surgery and decreases thereafter.

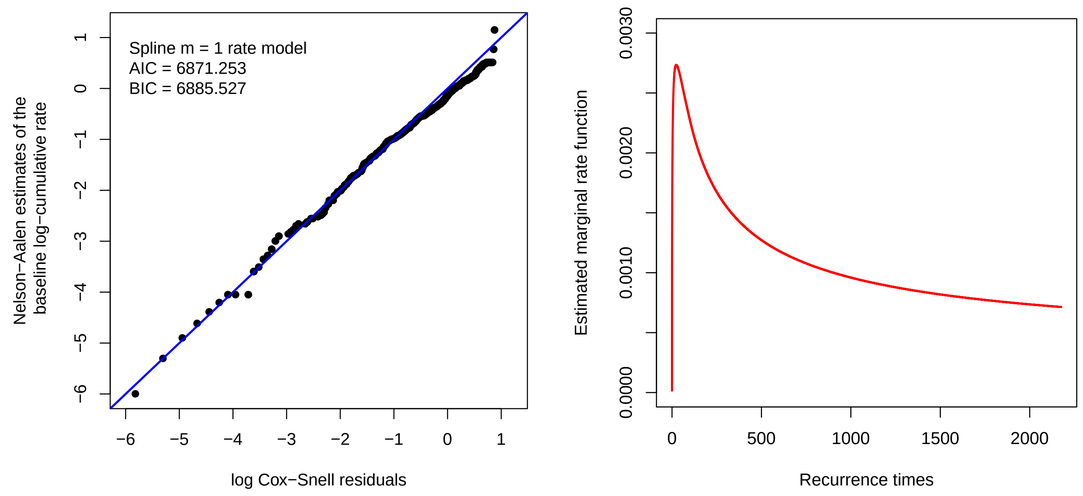

Figure 3: Cox-Snell residuals (left) and estimated marginal rate function (right) of the proposed model with $m = 1$ knot.

Conclusion and further work

- ✓ The proposed model is innovative in the sense that a RCS function is used to deduce the conditional distribution of the gap times between recurrent events;
- ✓ A zero-recurrence proportion is also incorporated to conveniently take into account the existence of subjects that will never experience any recurrence.
- ✓ In the application to the re-hospitalization data, the new model revealed to be very flexible, only requiring 1 internal knot to have an excellent fit to the data;
- ✓ For future research it would be interesting to include a random effect term in order to deal with the unobserved heterogeneity across subjects (which originates a frailty model).

Acknowledgments

This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, under the projects UIDB/00006/2020 (CEAUL – Centro de Estatística e Aplicações) and UIDB/04674/2020 (Center for Research in Mathematics and Applications (CIMA) related with the Statistics, Stochastic Processes and Applications (SSPA) group). I. Sousa-Ferreira also acknowledges FCT for the PhD grant DFA/BD/6459/2020.

References

[1] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

[2] Royston, P., and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, **21**(15), 2175 – 2197.

[3] Zhao, X., and Zhou, X. (2012). Modeling gap times between recurrent events by marginal rate function. *Computational Statistics & Data analysis*, **56**(2), 370 – 383.

Ciência 2021 – Encontro com a Ciência e Tecnologia em Portugal, June 28–30, 2021.