Parametric regression models for recurrent events analysis based on Chen distribution

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

[1] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal and Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

[2] Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal e Centro de Investigação em Matemática e Aplicações, Portugal

E-mail for correspondence: ivo.ferreira@staff.uma.pt

Abstract: In this paper, two parametric regression models based on Chen distribution are proposed for situations where recurrent events have the same or different risks of occurrence. Inference is based on a maximum likelihood approach, which ensures consistent parameter estimators. However, since in recurrent event data there is within-subject correlation, the "one step" jackknife estimator is used. An application on a real data set is also provided for illustrative purposes.

Keywords: Bathtub-shaped hazard; Chen distribution; Parametric survival models; Recurrent events; Variance corrected models.

1 Introduction

Recurrent events per subject are frequently observed in longitudinal studies. This kind of data arises in a wide variety of situations, such as medical studies on disease recurrence. Several approaches for modelling time to recurrent events were suggested (Cook and Lawless, 2007), being the extensions of the semi-parametric Cox model the most applied. Another approach is based on fully-parametric modelling the time, which remains under-worked in the recurrent events setting, possibly because semi-parametric models often tend to be preferred due to their weaker assumptions. However, when they are adequate, parametric models are able to estimate the regression coefficients more efficiently.
In the parametric approach, practitioners commonly select a distribution among those that they understand well, without considering other possibilities that may

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

better describe the data. With the intention of evaluating the appropriateness of other less known distributions with interesting properties, we propose two parametric regression models based on the distribution presented by Chen (2000).

2 Methodology

Chen (2000) proposed a two-parameter lifetime distribution in which the hazard function at time t is given by

$$h(t) = \lambda\gamma t^{\gamma-1}\exp(t^{\gamma}), \quad t > 0, \tag{1}$$

where $\lambda > 0$ and $\gamma > 0$ are the scale and shape parameters, respectively. Since $dh(t)/dt = (\gamma(t^{\gamma}+1)-1)h(t)/t$, it can be seen that $h(t)$ is: i) bathtub-shaped when $\gamma < 1$ (decreasing for $0 \leq t \leq (1/\gamma - 1)^{1/\gamma}$ and increasing for $t > (1/\gamma - 1)^{1/\gamma}$); and ii) monotonically increasing when $\gamma \geq 1$.

The two proposed models are based on the gap time modelling. Suppose that there are n independent subjects in study and that each one can experience a maximum of K_i $(i = 1, \ldots, n)$ recurrences of an event. For the ith subject, let the random variable T_{ik} be the time since the beginning of the study until the occurrence of the kth event $(k = 1, \ldots, K_i)$. The time intervals between two consecutive events (gap times) of the ith subject are defined as $Y_{ik} = T_{ik} - T_{i,k-1}$, $k = 1, \ldots, K_i$, where $0 \equiv T_{i0} < T_{i1} < \ldots < T_{iK_i}$. Furthermore, the observations are subjected to a right-censoring mechanism and the censoring is assumed to be non-informative.

For the formulation of the two new regression models, firstly we consider that the risk of occurrence of an event is not affected by previous events. Reparametrizing the scale parameter in (1) as $\lambda\exp(\boldsymbol{\beta}'\boldsymbol{z}_i)$ a proportional hazards model is obtained, where $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ is a vector of regression parameters and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})'$ is a vector of covariates. Thus, for the ith subject the model is given by

$$h(y_{ik}; \boldsymbol{z}_i) = \lambda\gamma y_{ik}^{\gamma-1}\exp(y_{ik}^{\gamma} + \boldsymbol{\beta}'\boldsymbol{z}_i), \tag{2}$$

where y_{ik} is the gap time of the ith subject with respect to the kth event. Secondly, a more complex situation is considered wherein repeated events have different risks of occurrence. Hence, the gap times are modelled through an event-specific baseline hazard function with the form (1) but with event-specific parameters. Thus, for the ith subject and the kth event, the hazard function is

$$h(y_{ik}; \boldsymbol{z}_i) = \lambda_k\gamma_k y_{ik}^{\gamma_k-1}\exp(y_{ik}^{\gamma_k} + \boldsymbol{\beta}'\boldsymbol{z}_i), \tag{3}$$

where y_{ik} is defined as previously, $\lambda_k\exp(\boldsymbol{\beta}'\boldsymbol{z}_i)$ is the event-specific scale parameter and γ_k is the event-specific shape parameter for the kth event. Notice that, for the kth event, the hazards are proportional. In general, the models (2) and (3) are unstratified and stratified regression models, respectively.

Concerning the risk set definition, since in the unstratified model (2) all events share a common baseline hazard function, all the subjects' risk intervals contribute to the risk set of any given event, regardless of the number of events experienced by each subject. On the other hand, as the stratified model (3) has

an event-specific baseline hazard function, only the risk intervals of the subjects who have experienced $k-1$ events contribute to the risk set of the kth event. The inferential procedure is based on the maximum likelihood asymptotic theory, assuming independence between the gap times. This approach yields consistent and asymptotically normal parameter estimators, even in the presence of within-subject correlation due to the existence of more than one observation per subject. However, the usual estimator of the covariance matrix should not be considered. In fact, this is a naïve approach that usually deflates the standard error, resulting in smaller p-values and unduly optimistic results. Therefore, to obtain the variance estimates, the "one step" jackknife estimator proposed by Lipsitz *et al.* (1994) was also applied. This estimator is asymptotically equivalent to other well-known variance estimators, such as White's estimator. Moreover, it has the advantage of being easier to program with a simple loop, resulting in faster computation.

3 An application to CGD data

The chronic granulamatous disease (CGD) data is a widely used data set in recurrent event framework. The data represent the time (in days) to a serious infection in 128 patients, of whom 63 received gamma interferon (rIFN-g) and 65 received a placebo, for approximately a year. By the end of the study, 14, 5 and 1 subjects of the rIFN-g group had experienced 1, 2 and 3 events, respectively; while 30, 12, 7, 3, 2, 1 and 1 subjects of the placebo group had 1, 2, ..., 7 events, respectively. Since the number of subjects at risk gradually decreases, in the stratified model the last strata were agglomerated from the 4th event onwards to avoid obtaining unreliable estimates. The full data set is reported in Fleming and Harrington (1991).

Here, only 2 covariates were included in the models: treatment (rIFN-g or placebo) and age (in years). Besides fitting the unstratified (2) and stratified (3) models, the results of the time to 1st event model are also presented (see Table 1). In the time to 1st event model, the usual ($\widehat{SE}$) and robust ($\widehat{SE}_r$) standard error estimates agree closely, whereas in the models for recurrent events $\widehat{SE}_r$ is in general larger than $\widehat{SE}$, as anticipated. When a very inflated $\widehat{SE}_r$ is obtained it means that there is less variation within subjects than between, while a very deflated $\widehat{SE}_r$ means that there is less variation between subjects than within. Both situations constitute evidence of violation of the independence assumption. The second situation only happened in the stratified model, more precisely in the standard errors of $\widehat{\lambda}_4$ and $\widehat{\gamma}_4$. This is not surprising for the later strata since the risk set decreases with the recurrences, making the resulting group less heterogeneous.

The time to 1st event and unstratified models yielded nearly identical regression estimates. Comparing these two models, for the treatment effect (β_1) there is an apparent reduction of 20% in $\widehat{SE}$, while using $\widehat{SE}_r$ the correct and more realistic reduction is of 1%. Applying the robust Wald test, all the models revealed that β_1 is statistically significant at the 0.01 level, whereas the age effect (β_2) is not. Note that, for the unstratified model, if the $\widehat{SE}$ is considered the p-value associated with β_2 will be 0.024. It is also relevant to notice that the estimates of λ and γ of the time to 1st event model and the estimates of λ_1 and γ_1 of the stratified

TABLE 1. Parameter estimates of each model for the CGD data.

Parameter		Estimate	$\widehat{SE}$	$\widehat{SE}_r$	95% CI	p-value
Time to 1st event model	λ	0.007	0.004	0.004	(0.000, 0.015)	< 2e-16
	γ	0.283	0.017	0.019	(0.246, 0.320)	< 2e-16
	β_1	-1.078	0.328	0.333	(-1.731, -0.426)	1.20e-03
	β_2	-0.027	0.017	0.017	(-0.061, 0.006)	0.107
Unstratified model	λ	0.015	0.005	0.006	(0.003, 0.027)	< 2e-16
	γ	0.264	0.013	0.012	(0.240, 0.287)	< 2e-16
	β_1	-1.083	0.262	0.330	(-1.730, -0.436)	1.04e-03
	β_2	-0.029	0.013	0.015	(-0.059, 0.000)	0.053
Stratified model	λ_1	0.006	0.003	0.004	(0.000, 0.013)	< 2e-16
	λ_2	0.019	0.011	0.009	(0.001, 0.037)	< 2e-16
	λ_3	0.012	0.013	0.016	(0.000, 0.043)	< 2e-16
	λ_4	0.012	0.013	0.008	(0.000, 0.028)	< 2e-16
	γ_1	0.282	0.017	0.019	(0.244, 0.320)	< 2e-16
	γ_2	0.268	0.026	0.026	(0.217, 0.320)	< 2e-16
	γ_3	0.307	0.041	0.064	(0.182, 0.432)	< 2e-16
	γ_4	0.345	0.054	0.048	(0.250, 0.440)	< 2e-16
	β_1	-0.845	0.272	0.308	(-1.449, -0.241)	6.12e-03
	β_2	-0.023	0.013	0.013	(-0.048, 0.002)	0.069

model are practically the same, although that doesn't happen with the covariates effect.

For all models, the estimates of γ are less than 1 indicating that the hazard function is bathtub-shaped. In addition, the adequacy of the Chen distribution to CGD data was informally evaluated through a plot where, for the first 3 events, the Kaplan-Meier and model-based estimates of the survival function were depicted (see Figure 1). The curves showed close agreement, indicating that this distribution is also a suitable parametric alternative for modelling the time to recurrent infections.

The clinicians conducting the CGD study suggested that the risk of a subsequent event remained unchanged regardless of the number of events observed (Fleming and Harrington, 1991), suggesting the use of the unstratified model. However, the AIC values of the unstratified and stratified models were 1079.85 and 1071.22, respectively, pointing to the stratified model as the best choice. Actually, in Figure 1 it can be observed that the survival curves from the second event are relatively close, but clearly different from the first event. This could be the main reason why the stratified model has a slightly smaller AIC value.

The computational implementation was developed in R software (R Core Team, 2019), version 3.5.3, where the maximum likelihood estimates were obtained using the Newton-Raphson maximization procedure. The Broyden-Fletcher-Goldfarb-Shanno iterative method was also used as an alternative procedure, allowing to confirm the obtained results and revealing fewer problems of convergence.

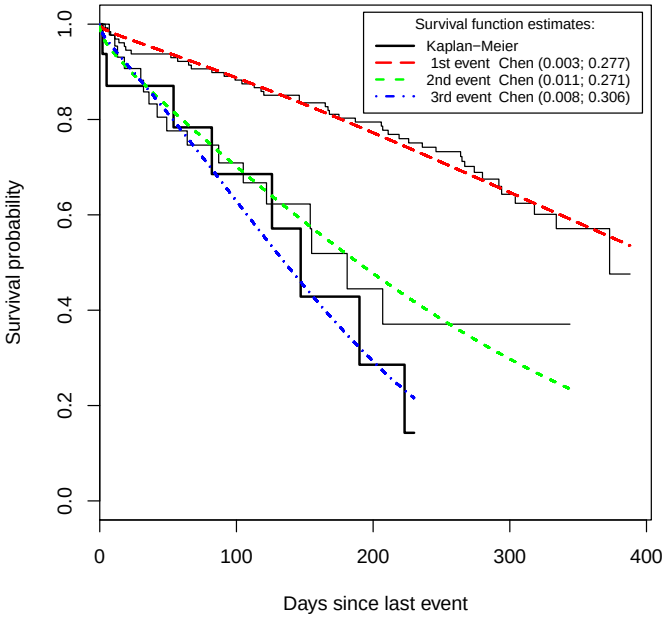

FIGURE 1. Estimated survival functions for the first 3 events, based on the Kaplan-Meier estimator and the Chen($\lambda;\gamma$) distribution, where λ and γ are the scale and shape parameters, respectively.

4 Conclusion and further work

In this paper, two regression models based on Chen distribution are proposed to analyse gap times between recurrent events, whether they have the same (unstratified model) or distinct (stratified model) risks of occurrence. As far as we know, this is the first time that the Chen distribution is used to formulate a regression model for recurrent events as well as a bathtub-shaped hazard function is considered in this context.

In general, for CGD data, the results reflects a satisfactory fit of the proposed models. Moreover, the inflated robust standard error estimates highlighted the importance of taking into account the within-subject correlation. In fact, this is a marginal method in which the parameters are estimated ignoring correlation, followed by a correction of the usual variance estimate. In other words, these are variance corrected models.

For future research it would be interesting to apply a conditional method to these two models, wherein the correlation is modelled using a random effect (which originates a frailty model).

Acknowledgments: I. Sousa-Ferreira is grateful to the *Universidade de Lisboa* for his PhD scholarship. The research was partially sponsored by portuguese funds through *FCT – Fundação para a Ciência e a Tecnologia*, under the projects

UID/MAT/00006/2019 (*Centro de Estatística e Aplicações*) and UID/MAT/04674/2019 (*Centro de Investigação em Matemática e Aplicações*).

References

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters*, **49**, 155 – 161.

Cook, R. J. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media.

Fleming, T. R. and Harringon, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Lipsitz, S. R., Dear, K. B. G. and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics*, **50**, 842 – 846.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`