Modeling the cumulative incidence function of clustered competing risks data: a multinomial GLMM approach

Henrique Aparecido Laureano* Wagner Hugo Bonat*

April 22, 2021

Abstract

Clustered competing risks data are a complex failure time data scheme. Its main characteristics are the cluster structure, which implies a latent within-cluster dependence between its elements, and its multiple variables competing to be the one responsible for the occurrence of an event, the failure. To handle this kind of data, we propose a full likelihood approach, based on a generalized linear mixed model instead a usual complex frailty model. We model the competing causes in the probability scale, in terms of the cumulative incidence function (CIF). A multinomial distribution is assumed for the competing causes and censorship, conditioned on the latent effects. The latent effects are accommodated via a multivariate Gaussian distribution. The CIF is specified as the product of an instantaneous risk level function with a failure time trajectory level function. The estimation procedure is performed through the R package TMB (Template Model Builder), an `C++` based framework with efficient Laplace approximation and automatic differentiation routines. A large simulation study is performed, based on different latent structure formulations. The model presents to be of difficult estimation, with our results converging to a latent structure where the risk and failure time trajectory levels are correlated.

Keywords: Clustered competing risks; Within-cluster dependence; Multinomial generalized linear mixed model (GLMM); TMB: Template Model Builder; Laplace approximation; Automatic differentiation (AD).

1 Introduction

Failure time data is the branch of Statistics responsible to handle random variables describing the time until the occurrence of an event, a failure (Kalbfleisch and Prentice; 2002; Hougaard; 2000). The time until a failure is called survival experience, the modeling object. To accommodate the number of possible causes for a failure and the form how they relate, different failure time data layouts were proposed and are shown in Figure 1. The first two are special cases of the last, a multistate process. The special cases are characterized by the presence of only absorbent states, besides the initial state 0. A multistate process is characterized by the presence of at least one intermediate state. In this

*Laboratory of Statistics and Goeinformation, Departament of Statistics, Paraná Federal University, Curitiba, Brazil. E-mail: laureano@ufpr.br

work, our focus is on the competing risks process, more specifically, its clustered version i.e., with groups of elements sharing some non-observed latent dependence structure.

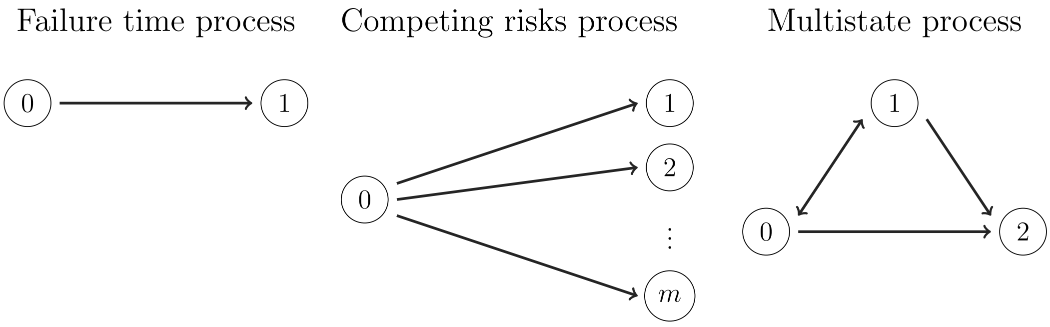

Figure 1: Illustration of failure time data layouts.

In the analysis of failure time data, the survival experiences are usually modeled in the hazard (failure rate) scale, and when the latent within-cluster dependence is accommodated we have the so-called frailty models (Prentice et al.; 1978; Clayton; 1978; Valpel et al.; 1979; Liang et al.; 1995; Petersen; 1998; Therneau and Grambsch; 2000). The use of frailty models implies complicated likelihood functions and its inference is done via elaborated and slow EM algorithms (Nielsen et al.; 1992; Klein; 1992) or inefficient MCMC schemes (Hougaard; 2000). Besides, its inference is attached to the modeling scale, i.e. the interpretations are in terms of hazard rates. A less usual scale but with a more appealing interpretation, is to model the survival experiences in the probability scale.

For competing risks data (Andersen et al.; 2012), the work on the probability scale is done by means of the cumulative incidence function (CIF), with the main modeling approach being the subdistribution (Fine and Gray; 1999). For clustered competing risks data there are also some available options but without any predominance. There are some complicated pairwise nonparametric approaches for bivariate CIFs (Cheng et al.; 2007, 2009), and semiparametric approachs based on composite likelihoods (Lindsay; 1988; Cox and Reid; 2004; Varin et al.; 2011) for also bivariate CIFs (Shih and Albert; 2009).

The class of generalized linear models (GLMs) (Nelder and Wedderburn; 1972) is probably the most popular statistical modelling framework. Despite its flexibility, the GLMs are not suitable for dependent data. In the case of longitudinal data, it is essential that the regression model take into account the longitudinal and/or grouped data structure. According to Diggle et al. (2002) longitudinal data are repeated measures evaluated on the same subjects over time, that are potentially correlated. Dependent data can also arises in studies with block designs, spatial and multilevel data (Verbeke and Molenberghs; 2001; Fitzmaurice et al.; 2008). For the analysis of such data several methods have been proposed over the last four decades.

Laird and Ware (1982) proposed the random effects regression models for longitudinal data analysis. Breslow and Clayton (1993) presented the generalized linear mixed models (GLMMs) for the analysis of non-Gaussian outcomes. Masarotto and Varin (2012) developed a class of marginal models for modelling dependence structures in the analysis of longitudinal data, time series and spatial based on Gaussian copula models.

The main goal of this study is to propose the . In this paper, we will investigate the as an alternative to . R (R Core Team; 2021) package `TMB` (Kristensen et al.; 2016).

The main contributions of this article are: (i) introducing the unit gamma distribution into the GLMMs framework; (ii) performing a extensive simulation study to check the properties of the the maximum likelihood estimator to deal with longitudinal continuous bounded outcomes; (iii) applying the proposed model in two data sets from different fields of application; (iv) providing R code and `C++` implementation for the unit gamma mixed models.

The work are organized as follows. Section 2, Section 3, Section 4. Finally, the main contributions of the article are discussed in Section 5.

2 multiGLMM: a multinomial GLMM for clustered competing risks data

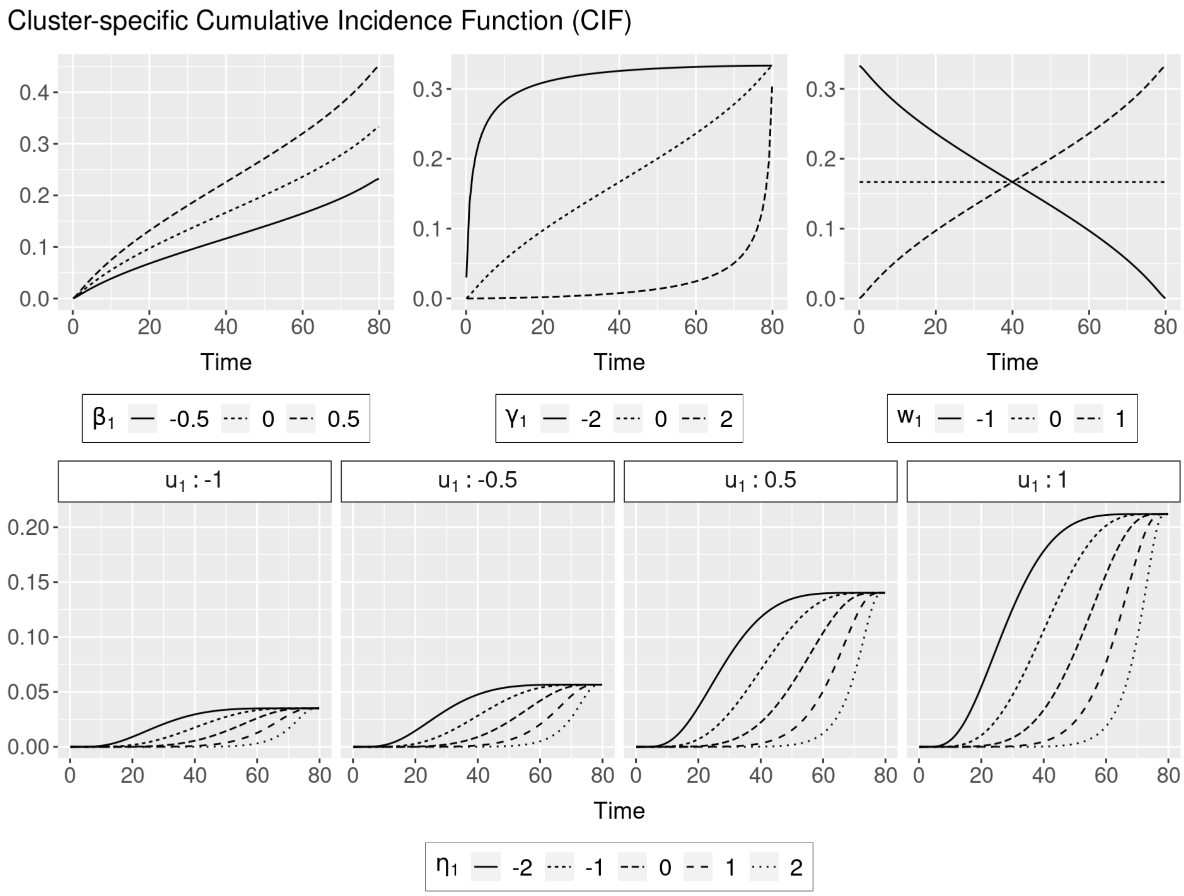

Figure 2: Curve behaviors for different parameter settings, showing then the corresponding parameter effects in a cluster-specific cumulative incidence function (CIF).

3 Estimation and inference

4 Simulation studies

5 Discussion

Supplementary material

References

Andersen, P. K., Geskus, R. B., de Witte, T. and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls, *International Journal of Epidemiology* **31**(1): 861–870.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American statistical Association* **88**(421): 9–25.

Cheng, Y., Fine, J. P. and Kosorok, M. R. J. (2007). Nonparametric Association Analysis of Bivariate Competing-Risks Data, *Journal of the American Statistical Association* **102**(480): 1407–1415.

Cheng, Y., Fine, J. P. and Kosorok, M. R. J. (2009). Nonparametric Association Analysis of Exchangeable Clustered Competing Risks Data, *Biometrics* **65**(1): 385–393.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial rendency in chronic disease incidence, *Biometrika* **65**(1): 141–151.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities, *Biometrika* **91**(3): 729–737.

Diggle, P., Heagerty, P., Liang, K. Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*, second Edition edn, Oxford University Press, United Kingdom.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards models for the subdistribution of a competing risk, *Journal of the American Statistical Association* **94**(446): 496–509.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal Data Analysis*, CRC Press.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, second Edition edn, John Wiley & Sons, Inc., Hoboken, New Jersey.

Klein, J. P. (1992). Semiparametric estimation of random effects using cox model based on the em algorithm, *Biometrics* **48**(1): 795–806.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J. and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation, *Journal of Statistical Software* **70**(5): 1–21.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**(4): 963–974.

Liang, K. Y., Self, S., Bandeen-Roche, K. J. and Zeger, S. L. (1995). Some recent developments for regression analysis of multivariate failure time data, *Lifetime Data Analysis* **1**(1): 403–415.

Lindsay, B. G. (1988). Composite likelihood methods, *Comtemporary Mathematics* **80**(1): 221–239.

Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression, *Electronic Journal of Statistics* **6**(1): 1517–1549.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**(3): 370–384.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. (1992). A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models, *Scandinavian Journal of Statistics* **19**(1): 25–43.

Petersen, J. H. (1998). An Additive Frailty Model for Correlated Life Times, *Biometrics* **54**(1): 646–661.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks, *Biometrics* **1**(1): 541–554.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Shih, J. H. and Albert, P. S. (2009). Modeling Familial Association of Ages at Onset of Disease in the Presence of Competing Risk, *Biometrics* **66**(4): 1012–1023.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York.

Valpel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in Individual Frailty on the Dynamics of Mortality, *Demography* **16**(1): 439–454.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods, *Statistica Sinica* **21**(1): 5–42.

Verbeke, G. and Molenberghs, G. (2001). *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, New York.