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

Henrique Aparecido Laureano * Wagner Hugo Bonat*

April 15, 2021

Abstract

Clustered competing risks data is a special case of failure time data. Besides the cluster structure which implies a latent within-cluster dependence between its elements, this kind of data is characterized by 1) multiple causes/variables competing to be the one responsible for the occurrence of an event, a failure; and 2) censorship, when the event of interest happens or not for none of the competing causes, in the study period. To handle this type of data, we propose a generalized linear mixed model (GLMM) i.e., a latent-effects framework, instead of a usual survival model. In survival analysis, the modeling is usually done by means of the hazard rate, and the within-cluster dependence accommodation ends by generating a complicated likelihood function, sometimes intractable. We, on the other hand, model the clustered competing causes in the probability scale, in terms of the cumulative incidence function (CIF) of each competing cause. In our framework, we suppose a multinomial probability distribution for the competing causes and censorship, conditioned on the latent effects. The latent effects are accommodated via a multivariate Gaussian distribution and are modeled by the parameters of its covariance matrix. The probability distributions are connected via CIF, modeled here following specification, based on its decomposition as the product of an instantaneous risk level function with a trajectory time level function. The latent effects are inserted in those level functions. To make the model

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

parameters estimation the most efficient as possible, we use the template model builder (TMB) . With this `R` package, we have 1) the log-likelihood function written in `C++`; 2) access to efficient linear algebra libraries; 3) efficient Laplace approximation implementation for the latent-effects; and 4) an automatic differentiation (AD) routine, the state-of-the-art in derivatives computation. To check the estimability of our model a large simulation study is performed, based on different latent structure formulations, with the aim to verify which one is most adequate to real scenarios. The model presents to be of difficult estimation, with our results converging to a latent structure where the risk and trajectory time levels are correlated. In scenarios with high CIF the model exhibits the better results, but still with an excessive variance, showing that improvements are necessary.

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

1 Introduction

Regression models are the main statistical tool for investigating the relationship between a response variable and a set of explanatory variables. The class of generalized linear models (GLMs) (Nelder and Wedderburn; 1972) is probably the most popular statistical modelling framework to deal with Gaussian and non-Gaussian outcomes. Despite its flexibility, the GLMs are not suitable for response variables with support limited to the interval $(0, 1)$. In general, continuous bounded variables appear in the form of rates, proportions, indexes and percentages and they can be used in many research areas.

The analysis of bounded variables is generally performed by the beta () and simplex () regression models. Besides that, other regression models were proposed to analyze continuous bounded variables on the interval $(0, 1)$. Some examples are the unit-Weibull (), Johnson S_B (), Kumaraswamy () and unit gamma () regression models. Additionally, using second-moment assumptions developed a flexible class of regression models to deal with continuous bounded variables on the interval $[0, 1]$.

Although these models are useful in many applications, they are usually limited to analyze independent 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 longitudinal data are repeated measures evaluated on the same subjects over time, that are potentially cor-

related. Dependent data can also arises in studies with block designs, spatial and multilevel data (). For the analysis of such data several methods have been proposed over the last four decades.

proposed the random effects regression models for longitudinal data analysis. presented the generalized linear mixed models (GLMMs) for the analysis of non-Gaussian outcomes. and extended the GLMs for the analysis of longitudinal data using a generalized estimating equation (GEE) approach. 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.

Based on the aforementioned approaches, some regression models have been proposed to deal with longitudinal continuous bounded outcomes. GLMMs based on beta distribution were employed in medical research (), social sciences () and behavioral studies (). Other regression models based on the simplex distribution were proposed for modelling longitudinal data (). Under the likelihood paradigm, the simplex mixed models with applications is discussed in .

The main goal of this study is to propose the unit gamma mixed model to deal with longitudinal continuous bounded outcomes. The unit gamma distribution is new in the literature and has been explored in other contexts, like control charts (), comparison between different methods for parameter estimation () and likelihood ratio tests (). In this paper, we will investigate the unit gamma distribution as an alternative to beta distributions for the analysis of dependent data bounded on the interval $(0, 1)$. We considered this distribution into the GLMM framework in order to fit regression models with random effects. We use automatic differentiation () and Laplace approximation (Tierney and Kadane; 1986) for efficient estimation of the proposed model through the `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 **??** presents the unit gamma mixed models. Section **??** describes the method proposed for parameter estimation and inference. The results of simulation studies are reported in Section **??**. Section **??** ilustrates the application of the model in two data sets. Finally, the main contributions of the article are discussed in Section **??**.

2 Unit gamma mixed models

3 Estimation and inference

4 Simulation studies

Figure 1: Fitted values by quarters, locations, random intercept.

5 Discussion

Supplementary material

References

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.

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.

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

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**(393): 82–86.