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 (Ferrari and Cribari-Neto; 2004) and simplex (Barndorff-Nielsen and Jørgensen; 1991) 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 (Mazucheli et al.; 2020), Johnson S_B (Lemonte and Bazán; 2016), Kumaraswamy (Mitnik and Baek; 2013) and unit gamma (Mousa et al.; 2016) regression models. Additionally, using second-moment assumptions Bonat et al. (2019) 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 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 et al. (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. Liang and Zeger (1986) and Zeger et al. (1988) extended the GLMs for the analysis of longitudinal data using a generalized estimating equation (GEE) approach. Masarotto et al. (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 (Hunger et al.; 2012), social sciences (Bonat, Ribeiro Jr and Zeviani; 2015; Bonat, Ribeiro Jr and Shimakura; 2015) and behavioral studies (Verkuilen and Smithson; 2012). Other regression models based on the simplex distribution were proposed for modelling longitudinal data (Song and Tan; 2000; Song et al.; 2004; Qiu et al.; 2008). Under the likelihood paradigm, the simplex mixed models with applications is discussed in Bonat et al. (2018).

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 (Lee Ho et al.; 2019), comparison between different methods for parameter estimation (Dey et al.; 2019) and likelihood ratio tests (Guedes et al.; 2020). 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 (Griewank and Walther; 2008) and Laplace approximation (Tierney and Kadane; 1986) for efficient estimation of the proposed model through the `R` (R Core Team; 2019) 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) pro-

viding `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

$$\ell(\boldsymbol{\theta}) = \frac{q}{2}\log(2\pi) - \frac{1}{2}\det\big(\mathcal{H}(\boldsymbol{\theta})\big) + \phi\log(z) - \log(\Gamma(\phi)) + (z-1)\log(y) + (\phi-1)\log(-\log(y)) - \frac{q}{2}\log(2\pi) - \frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{2\tau^2},$$

4 Simulation studies

4.1 Unit gamma mixed model with varying precision

5 Applications

5.1 Body fat percentage data set

$$\text{logit}(\mu_{ijk}) = \beta_0 + \beta_1\,\texttt{age}_i + \beta_2\,\texttt{BMI}_i + \beta_3\,\texttt{sex}_i + \beta_{4k}\,\texttt{IPAQ}_{ki} + \beta_{5j}\,\texttt{regions}_{ji} + u_{i1},$$

and

$$\log(\phi_{ij}) = \gamma_0 + \gamma_1\,\texttt{BMI}_i + \gamma_{2j}\,\texttt{regions}_{ji} + u_{i2}.$$

$$\begin{gathered} Y_{ijk}|u_i, u_{i,j} \sim \text{UG}(\mu_{ijk}, \phi), \\ \text{logit}(\mu_{ijk}) = \beta_0 + \beta_{1j}\texttt{quarter}_{ij} + \beta_{2k}\texttt{location}_{ik} + u_i + u_{i,j}, \\ u_i \sim \mathcal{N}(0, \tau_1^2) \;\text{ and }\; u_{i,j} \sim \mathcal{N}(0, \tau_2^2), \end{gathered}$$

Figure 1: Fitted values by quarters, locations, random intercept[1] and nested models[2].

6 Discussion

Supplementary material

References

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex, *Journal of Multivariate Analysis* **39**(1): 106–116.

Bonat, W. H., Lopes, J. E., Shimakura, S. E. and Ribeiro Jr, P. J. (2018). Likelihood analysis for a class of simplex mixed models, *Chilean Journal of Statistics* **9**(2).

Bonat, W. H., Petterle, R. R., Hinde, J. and Demétrio, C. G. (2019). Flexible quasi-beta regression models for continuous bounded data, *Statistical Modelling* **19**(6): 617–633.

Bonat, W. H., Ribeiro Jr, P. J. and Shimakura, S. E. (2015). Bayesian analysis for a class of beta mixed models, *Chilean Journal of Statistics* **6**(1): 3–13.

Bonat, W. H., Ribeiro Jr, P. J. and Zeviani, W. M. (2015). Likelihood analysis for a class of beta mixed models, *Journal of Applied Statistics* **42**(2): 252–266.

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.

Dey, S., Menezes, A. F. and Mazucheli, J. (2019). Comparison of estimation methods for unit-gamma distribution, *Journal of Data Science* **17**(4): 768–801.

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

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics* **31**(7): 799–815.

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

Griewank, A. and Walther, A. (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, Society for Industrial and Applied Mathematics (SIAM).

Guedes, A. C., Cribari-Neto, F. and Espinheira, P. L. (2020). Modified likelihood ratio tests for unit gamma regressions, *Journal of Applied Statistics* **47**(9): 1562–1586.

Hunger, M., Döring, A. and Holle, R. (2012). Longitudinal beta regression models for analyzing health-related quality of life scores over time, *BMC medical research methodology* **12**(1): 144.

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

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

Lee Ho, L., Fernandes, F. H. and Bourguignon, M. (2019). Control charts to monitor rates and proportions, *Quality and Reliability Engineering International* **35**(1): 74–83.

Lemonte, A. J. and Bazán, J. L. (2016). New class of Johnson SB distributions and its associated regression model for rates and proportions, *Biometrical Journal* **58**(4): 727–746.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22.

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

Mazucheli, J., Menezes, A., Fernandes, L., de Oliveira, R. and Ghitany, M. (2020). The unit-weibull distribution as an alternative to the kumaraswamy distribution for the modeling of quantiles conditional on covariates, *Journal of Applied Statistics* **47**(6): 954–974.

Mitnik, P. A. and Baek, S. (2013). The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation, *Statistical Papers* **54**(1): 177–192.

Mousa, A. M., El-Sheikh, A. A. and Abdel-Fattah, M. A. (2016). A gamma regression for bounded continuous variables, *Advances and Applications in Statistics* **49**(4): 305.

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

Qiu, Z., Song, P. X.-K. and Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data, *Scandinavian Journal of Statistics* **35**(4): 577–596.

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/*

Song, P. X.-K., Qiu, Z. and Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data, *Biometrical Journal* **46**(5): 540–553.

Song, P. X.-K. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data, *Biometrics* **56**(2): 496–502.

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

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

Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution, *Journal of Educational and Behavioral Statistics* **37**(1): 82–113.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach, *Biometrics* **44**(4): 1049–1060.