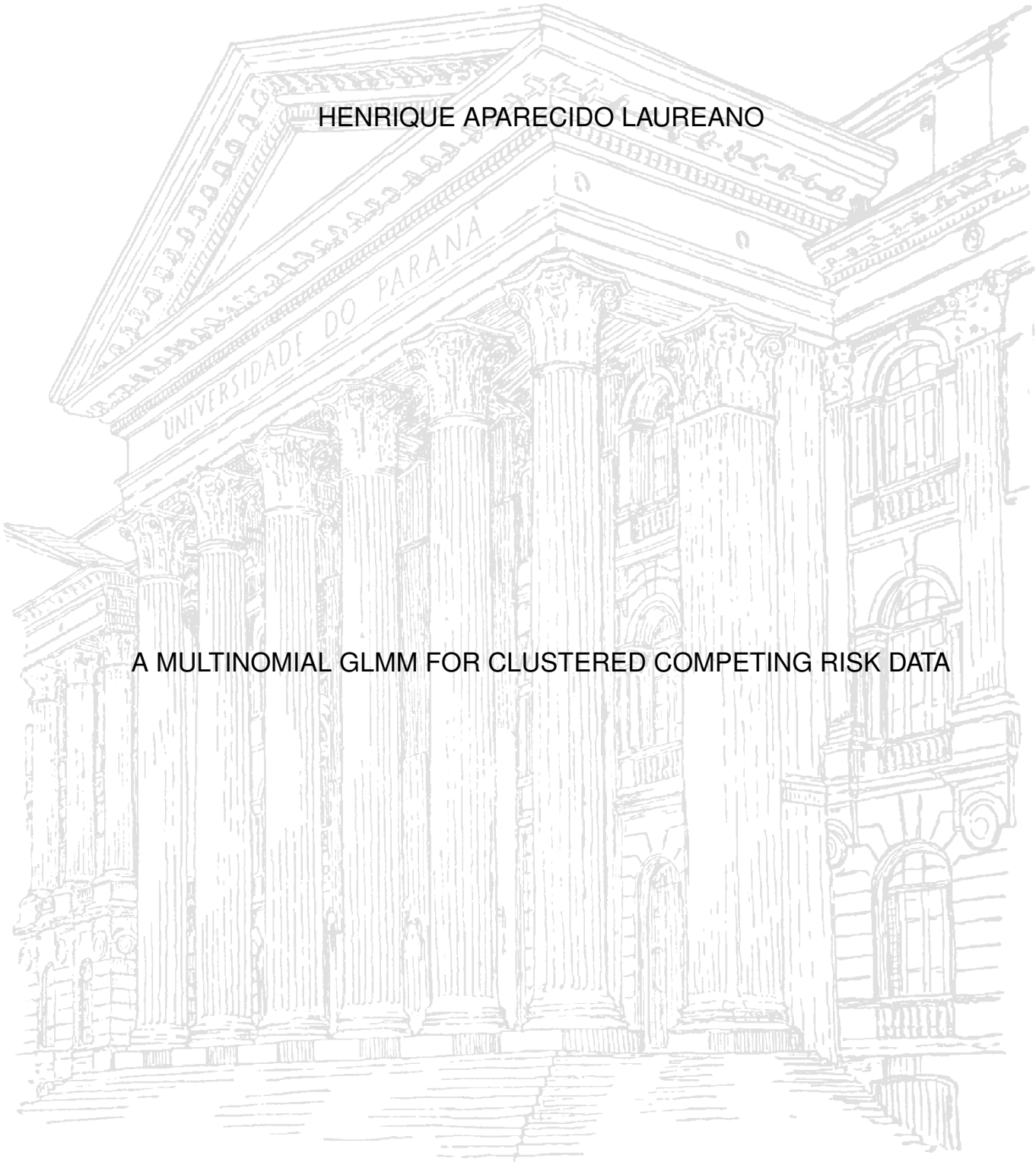
FEDERAL UNIVERSITY OF PARANÁ

HENRIQUE APARECIDO LAUREANO



A MULTINOMIAL GLMM FOR CLUSTERED COMPETING RISK DATA

CURITIBA

2020

HENRIQUE APARECIDO LAUREANO

A MULTINOMIAL GLMM FOR CLUSTERED COMPETING RISK DATA

Thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods Applied in Engineering, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Supervisor: Prof. PhD Wagner Hugo Bonat

Co-supervisor: Prof. PhD Paulo Justiniano Ribeiro Jr

CURITIBA

2020

HENRIQUE APARECIDO LAUREANO

A MULTINOMIAL GLMM FOR CLUSTERED COMPETING RISK DATA

Thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods Applied in Engineering, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Master thesis approved. XXX XX, 2020.

Prof. PhD Wagner Hugo Bonat
Supervisor

Prof. PhD Paulo Justiniano Ribeiro Jr
Co-supervisor

Prof. PhD ...
Internal Examinator - PPGMNE

Prof. PhD ...
Internal Examinator - PPGMNE

Prof. PhD ...
External Examiner -

CURITIBA
2020

To Celita and Olivio

ACKNOWLEDGEMENTS

As Moro said once, I'm thankful for everything and everyone.

"It's not supposed to be easy."
(Gregg Popovich)

ABSTRACT

Failure time data …

Keywords: Competing risks.

RESUMO

Dados de tempos de falha …

Palavras-chave: Riscos competitivos.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\mathbb{E}(\cdot)$	The mathematical expectation of a random variable $\cdot$

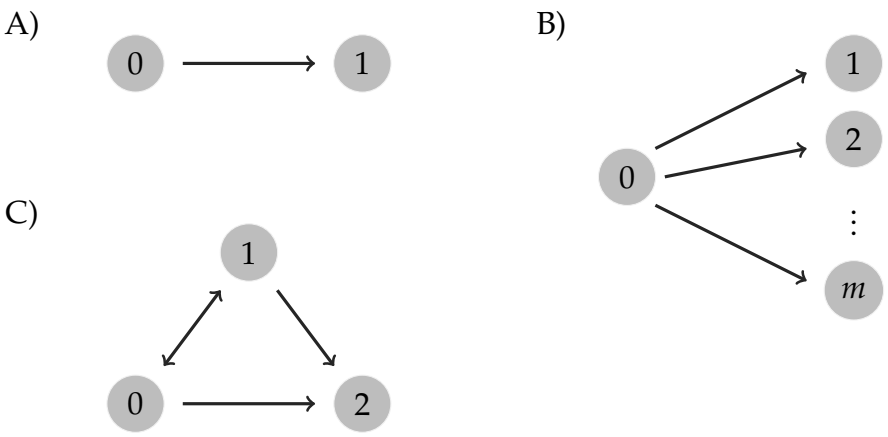
CONTENTS

1 INTRODUCTION

Consider a cluster of random variables. Each random variable represents the time until some event occurs. The random variables that compose the cluster are assumed to be correlated, i.e., the method to be used needs to be flexible enough to be able to accommodate that, possible, correlation. In this thesis, the cluster is a family, more precisely, part of a family - a pair of twins; the random variables are the time until the occurrence (or not) of an event in each twin; and the event under focus is the occurrence of cancer. The inspiration for this work came from Cederkvist et al. (2019), where they model the cause-specific cumulative incidence function of competing risks data allowing for within-cluster dependence of both risk and timing. What we do here is propose a simpler manner of doing the same, and easier to extend. But first, some definitions and theoretical contexts are welcome.

When the object under study is random variables representing the time until some event occurs, we're in the field of *failure time data* (KALBFLEISCH; PRENTICE, 2002). Such events are generally referred to as *failures*. Major areas of application are biomedical studies and industrial life testing. In this thesis, we maintain our focus on the former. In industrial life testing applications, is performed what is called a *reliability analysis*; in biomedical studies, is performed what is called *survival analysis*. Generally, the term survival analysis is applied when we're interested in the occurrence of only one event, a *failure time process*. When we're interested in the occurrence of more than one event, like now, we enter in the yard of *competing risks* and *multistate* models. A visual aid is presented on Figure 1 and a comprehensive reference is Kalbfleisch & Prentice (2002).

FIGURE 1 – BEHAVIOR ILLUSTRATIONS OF MULTISTATE MODELS FOR A) FAILURE TIME PROCESS; B) COMPETING RISKS; AND C) ILNESS-DEATH MODEL, THE SIMPLEST MULTISTATE MODEL



SOURCE: The author (2020).

Failure time and competing risk processes may be seen as particular cases of a multistate model. Besides the number of events (states) of interest, a big difference between a multistate model and its particular cases is that only in the multistate scenario we may have transient states, using a *stochastic process* language. In the particular cases, all the states, besides the initial state 0, are absorbents - once you reached it you don't leave. The simplest multistate model that exemplify this behavior is the so-called illness-death model, Figure 1 C). A patient enters the study (state 0) and it can get sick (state 1) or die (state 2); if sick it can recover (returns to state 0) or die. In this thesis, we'll work only with competing risk processes. For each individual, we have the time, age, until the occurrence (or not) of cancer.

When for some know or unknown reason we're able to see the occurrence of the event, we have what is called *censorship*. Still in the illness-death model: during the period of follow up the patient may not get sick or die, staying at state 0, this is called a *right-censorship*; The same for state 1. If a patient is in state 1 at the end of the study, we're *censored* to see him reaching the state 2 or returning to state 0. This is the inherent idea to censorship and must be present in the modeling framework.

In a survival model what we model is the survival experience. Usually, there are covariates (explanatory/independent variables) upon which failure time may depend. The model is defined by the *hazard* (failure rate), $\lambda(\cdot)$, at time t for an individual with covariate vector $\mathbf{X}_i$ and can be written as

$$\lambda(t\,;\,\mathbf{X}_i) = \lambda_0(t) \times c(\mathbf{X}_i\boldsymbol{\beta}), \tag{1.1}$$

where $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_p)$ is a vector of regression parameters; $\lambda_0(\cdot)$ is an arbitrary baseline hazard function; and c is a specific function form, that will depend on the chosen probability distribution for the failure time. The structure of equation 1.1 is made thinking in a simple failure time process, as in Figure 1 A). However, is easy to extend its idea. We basically have the equation 1.1 model for each cause-specific, in a competing risks process; or transition, in a multistate process. A complete and extensive detailing can be, again, found in Kalbfleisch & Prentice (2002).

In this thesis, we approach the case of the clustered competing risks. Besides the cause-specific structure, we have the fact that the disease occurrences are happening in related individuals (twins). This configures what is called *family studies*. We have a cluster (a family/pair of twins) dependence that needs to be considered. This, possible, dependence is something that we don't actually measure, but know (or just suppose) that exists. In the statistical modeling language, this type of characteristic receives the name of *random* or *latent* effect. A survival model with a latent effect, association, or unobserved heterogeneity, is called a *frailty model* (CLAYTON, 1978; VALPEL; MANTON; STALLARD, 1979). In its simplest form, a frailty is an unobserved random proportional-

ity factor that modifies the hazard function of an individual, or of related individuals. Frailty models are extensions of the equation 1.1 model.

Specifically in the competing risks setting, we have to choose which of two scales we want to work on. The hazard scale focusing on the cause-specific hazard or on the probability scale focusing on the cause-specific cumulative incidence. Both may complement each other (ANDERSEN et al., 2012). However, in family studies, there is often a strong interest in describing age at disease onset including within-family dependence. The distribution of age at disease onset is directly described by the cause-specific cumulative incidence. Therefore, the probability scale is the chosen one here.

Frailty models are generally defined based on the hazard scale and with a latent effect that modifies the hazard function via a proportionality factor. To work on the probability scale and with a latent effect that allows for within-cluster dependence of both risk and timing, Cederkvist et al. (2019) proposed a pairwise composite likelihood approach based based on a linear model with multinomial response distribution and multivariate normal latent effects (in a frailty model the common choice for the latent effects is the gamma distribution). The idea in this thesis is to try to do that in a simpler manner via a generalized linear mixed model (GLMM). Instead of concentrating on failure time data and consequently having a survival/frailty model on the hazard scale, or using a composite approach, we just build the joint likelihood function (multinomial distrtibution with a link function based on the cause-specific cumulative incidence function and appropriate latent effects), integrate out the latent effects and optimize the obtained distribution with respect to (wrt) its parameters. With this approach, we easily work on the probability scale and are able to specify the desired within-cluster dependence structure. To conclude, a brief introduction of a GLMM.

In a standard linear model we assume that the response variable, Y_i, conditioned on the covariates follows a normal distribution, and we model it's mean, $\mu_i \equiv \mathbb{E}(Y_i)$, via a linear combination. Generalizing this idea with a smooth monotonic "link function" g, we get a generalized linear model (GLM) (NELDER; WEDDERBURN, 1972) with the basic structure

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta},$$

where $\mathbf{X}_i$ is the i^{th} row of a model matrix $\mathbf{X}$, and $\boldsymbol{\beta}$ is a vector of unknown parameters. In a GLM the Y_i are independent and

$$Y_i \sim \text{some exponential family distribution.}$$

The *exponential family* of distributions includes many distributions that are useful for practical modelling, such as the Poisson (for counting data), binomial (dichotomic

data), gamma (continuous but positive) and normal (continuous data) distributions. The comprehensive reference for GLMs is McCullagh & Nelder (1989).

The insertion of a latent effect, $\mathbf{b}$, in that structure makes a mixed model. A linear model becomes a linear mixed model and a GLM becomes a generalized linear mixed model (GLMM). We now have

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\psi})$$

where the latent effect is assumed to follow a multivariate normal distribution of mean zero and a given variance-covariance structure.

1.1 GOALS

1.1.1 General goals

Propose a multinomial GLMM to the cause-specific cumulative incidence function of clustered competing risks data.

1.1.2 Specific goals

1. Simulate from the model following and adapting the guidelines from Cederkvist et al. (2019).
2. Write the model via Template Model Builder (TMB) (KRISTENSEN et al., 2016), possibly the most efficient way of doing so, and take advantage of its functionalities: compute all necessary gradients and Hessians via Automatic Differentiation and integrate out the latent effects of the joint likelihood via Laplace approximation.
3. Try different complexity levels of the proposed modeling framework to see how identifiable it is.
4. Apply the model to the Nordic Cancer Union (NCU) twins data.
5. Compare the results of the multinomial GLMM approach to the pairwise composite likelihood approach of Cederkvist et al. (2019).

1.2 JUSTIFICATION

In family studies examining disease occurrence in related individuals, key points of interest are the within-family dependence and determining the role of different risk factors. The within-family dependence may reflect both disease heritability and the

impact of shared environmental effects. The number of statistical models for competing risks data that accommodate the within-cluster (family) dependence is limited. We didn't find, e.g., any GLMM approach to do that and didn't find any justification to not do that. Therefore, we propose a multinomial GLMM apporach that accommodates these key points by modeling the cause-specific cumulative incidence function (that describes age at disease onset) of the competing risks, using a latent structure that allows the absolute risk and the failure time distribution to vary between clusters (here, families).

1.3 LIMITATION

This work restraint to the proposition and application of a multinomial model for competing risks data with a latent effect structure to accommodate within-cluster dependence with regard to both risk and timing. Given the elevated model complexity, hypothesis tests; residual analysis; and good-of-fit measures will not be approached.

1.4 THESIS ORGANIZATION

This thesis contains 6 chapters including this introduction. The chapter 2 presents a review of the main aspects of a general GLMM and its respective inference procedures. The chapter 3 presents the multinomial GLMM with its particular characteristics and in chapter 4 we describes how to simulate from the proposed model, and presents a dataset for a real application. In chapter 5 the obtained results are presented, and in chapter 6 we discuss the contributions of this thesis and present some suggestions for future work.

2 METHODS AND INFERENCE

3 MULTINOMIAL GENERALIZED LINEAD MIXED MODEL

4 DATASETS

5 RESULS

6 FINAL CONSIDERATIONS

6.1 FUTURE WORKS

BIBLIOGRAPHY

ANDERSEN, P. K.; GESKUS, R. B.; WITTE, T. de; PUTTER, H. Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, v. 31, n. 1, p. 861–870, 2012. Cited on page 14.

CEDERKVIST, L.; HOLST, K. K.; ANDERSEN, K. K.; SCHEIKE, T. H. Modeling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing. *Biostatistics*, v. 20, n. 2, p. 199–217, 2019. Cited 3 times on pages 12, 14, and 15.

CLAYTON, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial rendency in chronic disease incidence. *Biometrika*, v. 65, n. 1, p. 141–151, 1978. Cited on page 13.

KALBFLEISCH, J. D.; PRENTICE, R. L. *The Statistical Analysis of Failure Time Data*. Second edition. Hoboken, New Jersey: John Wiley & Sons, Inc., 2002. Cited 2 times on pages 12 and 13.

KRISTENSEN, K.; NIELSEN, A.; BERG, C. W.; SKAUG, H.; BELL, B. M. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, v. 70, n. 5, p. 1–21, 2016. Cited on page 15.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. Second edition. London: Chapman & Hall, 1989. Cited on page 15.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, v. 135, n. 3, p. 370–384, 1972. Cited on page 14.

VALPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, v. 16, n. 1, p. 439–454, 1979. Cited on page 13.