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

master thesis defense

Henrique Laureano (.github.io)
LEG @ UFPR

April 8, 2021

Clustered competing risk data

Key terms:

- Clustered: groups with a dependence structure (e.g. families);
- Causes competing by *something*.

Something?

- **Failure** of an industrial or electronic component;
- **Occurence** or **cure** of a disease or some biological process;
- **Progress** of a patient clinic state.

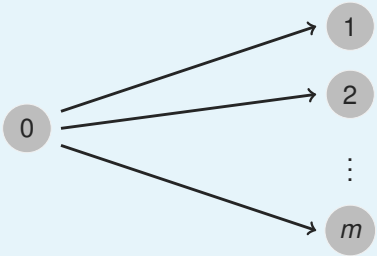
Independent of the application, always the same framework

Cluster	ID	Cause 1	Cause 2	Censorship	Time	Feature
1	1	Yes	No	No	10	A
1	2	No	No	Yes	8	A
2	1	No	No	Yes	7	B
2	2	No	Yes	No	5	A

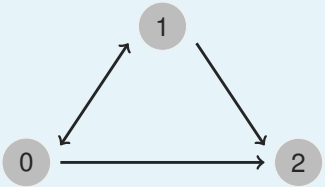
Big picture: Failure time data

Failure time process

Competing risk process



Multistate process

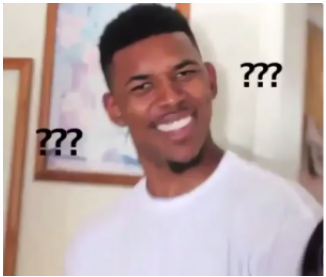


Same methodologies, different names.

Survival analysis Biomedical studies;
Reliability analysis Industrial life testing.

A comprehensive reference is Kalbfleisch and Prentice (2002)'s book.

Modeling clustered competing risks data



What?

Why?

How?

Failure time data $\rightarrow$ Survival models

First of all, we have to choose which scale we model the **survival experience**.

- Usually, is in the

$$\text{hazard (failure rate) scale}: \quad \lambda(t \mid \text{features}) = \lambda_0(t) \times c(\text{features}). \tag{1}$$

We have a Equation 1 for each competing cause.

The cluster dependence is something actually not measured...

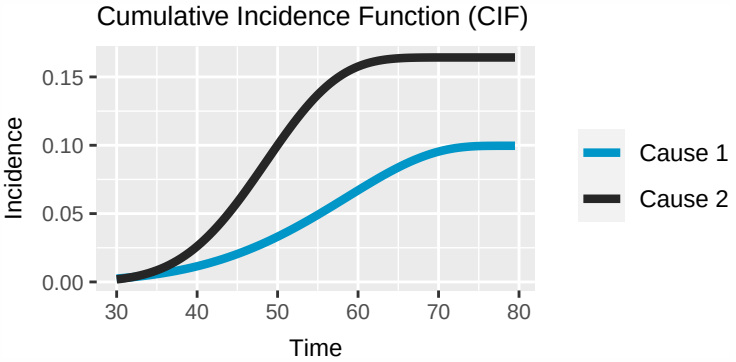
Not measured dependence $\rightarrow$ **random/latent effects** $\rightarrow$ Frailty models.

Full likelihood analysis with frailty models for competing risks data is generally complicated, when not impracticable.

- *Not* usually, the probability scale.

Probability scale $\rightarrow$ Cause-specific CIF

Besides the within-cluster dependence, there is an often interest in describing the time at event onset, directly described by the cause-specific



i.e., CIF $= \mathbb{P}[$ failure time $\leqslant t,$ a given cause $|$ features & latent effects $]$.

Formally,

for a cause-specific of failure k,
the cumulative incidence function (CIF) is defined as

$$
\begin{aligned}
F_k(t \mid \boldsymbol{x}) &= \mathbb{P}[T \leqslant t,\ K = k \mid \boldsymbol{x}] \\
&= \int_0^t f_k(z \mid \boldsymbol{x})\, \mathrm{d}z \quad (f_k(t \mid \boldsymbol{x}) \text{ is the (sub)density for the time to a type } k \text{ failure}) \\
&= \int_0^t \underbrace{\lambda_k(z \mid \boldsymbol{x})}_{\substack{\text{cause-specific} \\ \text{hazard function}}} \underbrace{S(z \mid \boldsymbol{x})}_{\substack{\text{overall} \\ \text{survival} \\ \text{function}}} \mathrm{d}z, \quad t > 0, \quad k = 1,\ \ldots,\ K.
\end{aligned}
$$


Again, a comprehensive reference is Kalbfleisch and Prentice (2002)'s book.


Here, we use the same CIF specification of Cederkvist et al. (2019).

Cederkvist et al. (2019)'s CIF specification

For two competing causes of failure,
the cause-specific CIFs are specified in the following manner

$$F_k(t \mid \boldsymbol{x},\ u_1,\ u_2,\ \eta_k) = \underbrace{\pi_k(\boldsymbol{x},\ u_1,\ u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[w_k g(t) - \boldsymbol{x}\gamma_k - \eta_k]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad t > 0, \quad k = 1,\ 2, \qquad (2)$$

i.e., as the product of a cluster-specific risk level with a cluster-specific failure time trajectory, resulting in a cluster-specific CIF.

That may reflect

- Disease **heritability**;
- The impact of shared **environmental effects**;
  - **Parental effects**: continuity of the phenotype across generations.

Our contribution: a hierarchical approach

Thinking on two competing causes

... for the outcome y_{ijt} of a subject i, family j, in the time t, we have

$$y_{ijt} \mid \underbrace{\{u_{1j}, u_{2j}, \eta_{1j}, \eta_{2j}\}}_{\text{latent effects}} \sim \text{Multinomial}(p_{1ijt}, p_{2ijt}, p_{3ijt})$$

$$\begin{bmatrix} u_{1j} \\ u_{2j} \\ \eta_{1j} \\ \eta_{2j} \end{bmatrix} \sim \text{Multivariate Normal} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_1} & \sigma_{u_1,u_2} & \sigma_{u_1,\eta_1} & \sigma_{u_1,\eta_2} \\ & \sigma^2_{u_2} & \sigma_{u_2,\eta_1} & \sigma_{u_2,\eta_2} \\ & & \sigma^2_{\eta_1} & \sigma_{\eta_1,\eta_2} \\ & & & \sigma^2_{\eta_2} \end{bmatrix} \right)$$

$$p_{kijt} = \frac{\partial \text{CIF}}{\partial t} = \frac{\partial}{\partial t} \underbrace{\pi_k(X, u_1, u_2 \mid \beta)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \underbrace{\Phi[w_k g(t) - X^\top \gamma_k - \eta_k]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}},$$

$k = 1, 2.$

Contributions & challenges

- A clear and simpler modeling structure;
- **There is no free lunch**
  Computational challenges overcame via an
  efficient implementation and estimation routines, the TMB;
- The data is very simple,
  we just know the outcome (**yes** or **no**);
- We have to be able to build the CIF curves;
- And accommodate the within-family dependence properly,
  that can happen in different manners;
- . . .

TMB: Template Model Builder

Quickly implement complex random effect models through simple C++ templates.
The R package combines

- CppAD: C++ automatic differentiation;
- Eigen: templated matrix-vector library;
- CHOLMOD: sparse matrix routines available from R;

to obtain an efficient implementation of the applied Laplace approximation with exact derivatives.

Also, key features are

- automatic sparseness detection;
- parallelism through BLAS;
- parallel user templates.

Thanks for watching and have a great day

Special thanks to

PPGMNE
Programa de Pós-Graduação em
Métodos Numéricos em Engenharia

CAPES

UFPR

Joint work with

Wagner H. Bonat
http://leg.ufpr.br/~wagner

Paulo Justiniano Ribeiro Jr.
http://leg.ufpr.br/~paulojus

henriquelaureano.github.io

References

Cederkvist, L., K. K. Holst, K. K. Andersen, and T. H. Scheike. 2019. “Modeling the Cumulative Incidence Function of Multivariate Competing Risks Data Allowing for Within-Cluster Dependence of Risk and Timing.” *Biostatistics* 20 (2): 199–217.

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