Revista Colombiana de Estadística - Applied Statistics
July 2023, volume 46, issue 2, pp. 145–161
http://doi.org/10.15446/rce.v46n2.98005

A Joint Model of Competing Risks in Discrete Time with Longitudinal Information

Un modelo de riesgos en competencia en tiempo discreto con información longitudinal

Adriana Marcela Salazar[a], Jaime Abel Huertas[b]

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

The survival competing risks model in discrete time based on multinomial logistic regression, proposed by Luo et al. (2016), models the non-linear and irregular shape of hazard functions by incorporating a time-dependent spline into the multinomial logistic regression. This model also directly includes longitudinal variables in the regression. Due to the issues arising from including both baseline and longitudinal covariates in the extended form as proposed, and considering that the latter may be subject to error, this article suggests an extension of the existing model. The proposed extension utilizes the concept of joint models for longitudinal and survival data, which is an effective approach for integrating simultaneousness both baseline and time-dependent covariates into the survival model.

***Key words*:** Discrete time; Joint model; Longitudinal model; Logistic regression; Survival model.

Resumen

El modelo de supervivencia de riesgos en competencia en tiempo discreto basado en regresión logística multinomial sugerida por Luo et al. (2016), modela la forma no lineal e irregular de las funciones de riesgo, incorporando un spline dependiente del tiempo en la regresión logística multinomial. Dicho modelo también incluye variables longitudinales directamente en la regresión. Debido a los problemas derivados de la inclusión tanto de covariables basales como longitudinales en la forma ampliada que hace la propuesta, y considerando que estas últimas pueden estar sujetas a error, este artículo sugiere una ampliación del modelo existente. La extensión propuesta utiliza

[a]Statistic. E-mail: amsalazar@unal.edu.co
[b]Ph.D. E-mail: jahuertasc@unal.edu.co

el concepto de modelos conjuntos para datos longitudinales y de supervivencia, que es un enfoque eficaz para integrar simultáneamente en el modelo de supervivencia tanto las covariables basales como las dependientes del tiempo.

***Palabras clave*:** Modelo conjunto; Modelo de supervivencia; Modelo longitudinal; Regresión logística; Tiempo discreto.

1. Introduction

Classification models, such as logistic regression, are useful for estimating the probability of an event based on other variables. We can also analyze the time to such an event by including it as a covariate in the model. However, to properly model event times, it is preferable to use survival models, such as the Cox model (Cox, 1972). The widespread use of logistic regression to model survival times began to be re-evaluated with the Cox model (Cox, 1972). An undoubtedly logical conclusion, since the logistic model does not consider the length of the follow-up time Myers et al. (1973). However, it wasn't until the early 1980s that mathematical comparisons emerged. Elandt-Johnson (1980) demonstrated how the likelihood of the logistic model approximates the partial likelihood of the Cox model when the event rate is low and follow-up time is not too long. Green & Symon (1983) showed that the estimation of the logistic regression model parameters is biased and loses precision as the follow-up time increases, and also concluded that: "In principle it would appear that the Cox model is superior to the logistic model since it utilizes time of response and hence incorporates more information." This discussion continued to evolve with studies such as Annesi et al. (1989), which reached similar conclusions as those already mentioned but generalizing to several qualitative and quantitative covariables.

When there are many tied survival times, the Cox model relies on approximations to estimate parameters, which can be computationally intensive and yield subpar results. On the other hand, it is worth noting that when the probabilities of survival events are low, both the Cox model and the logistic regression models produce similar outcomes. Therefore, when the probabilities of survival events are low and there are tied survival times, logistic regression can be a useful alternative for modeling time-to-events. This conclusion can be extended to the analysis of competing risks using multinomial logistic regression models, i.e., discrete-time survival models. A model of this type was proposed by Begg & Gray (1984) based on time-dependent covariates.

Luo et al. (2016) extended the multinomial regression model proposed by Begg & Gray (1984) by incorporating a time-dependent spline, which provides flexibility to the model when facing irregular risk behaviors and partially solves the loss of fit problem caused by long follow-up times. The purpose of Luo et al. (2016) was to estimate the probability of loan default at different time points, for credit risk assessment. The approach considers time-dependent covariates, treating each observed time point as a separate data line. The authors justified this data handling approach by noting that credit information is often recorded in this format. However, in practice, longitudinal data may also be available intermittently for each

subject and may be subject to errors. Furthermore, including both baseline covariates and longitudinal data in the model presents challenges because individuals with a longer credit history have a stronger influence on their baseline covariates than those who have recently started their credit, resulting in biased estimations. Because of this, we propose an extension to this model.

A useful method for integrating baseline and time-dependent covariates into the survival model is to jointly model longitudinal and survival data, where longitudinal information is summarized and incorporated into the survival model. In the credit risk context addressed by Luo et al. (2016), where the database also includes baseline covariates, the problem can be treated using joint modeling of longitudinal and competing risks data. Several proposals have been made in this area, such as those by Elashoff et al. (2007), Elashoff et al. (2008), Williamson et al. (2007), Li et al. (2009), or Teixeira et al. (2019). Since these proposals use semiparametric survival models, they also face the tie problem mentioned earlier. Hence, it is also reasonable to consider using multinomial regression in such cases, particularly when the probabilities of the survival events are low.

In this paper, we extend the important model proposed by Luo et al. (2016) for modelling competing risks in discrete time based on longitudinal information. We do this through the concept of joint modeling to analyze situations where it is necessary to include both baseline and longitudinal variables in the model. We specifically illustrate its application using a credit database. The subsequent sections, provide a concise overview of survival data analysis using logistic regression and the theory of joint modelling of longitudinal and survival data. Lastly, we present our proposal along with an application. This paper is based on the master's thesis of Salazar (2021).

2. Modeling Survival Data with Logistic Regression

After introducing the notation, we will provide a brief and clear explanation of the well-established relationship between the Cox model and logistic regression. This relationship is crucial as it justifies the use of logistic regression for survival analysis, and we will utilize it in our application. Next, we will introduce a multinomial regression model suitable for time-to-event analysis, for which we will provide an extension to enhance its capabilities.

Notation: In the multinomial model, we analyze a discrete time-to-event variable denoted as T, an event indicator δ, and a vector of independent covariates $\vec{X}$. In the model proposed by Luo et al. (2016), the covariates are considered time-dependent, hence the notation $\vec{X}(t)$.

The variable T represents the duration from the start date until the occurrence of an event, while δ indicates the different outcomes. If there are more than two possible outcomes, it suggests a competing risks scenario. In the case of two outcomes, we assume that the time to event T can be right-censored non-informatively by a censoring time C. Here, the observed data for each subject $i = 1, \ldots, n$ is denoted as (Y_i, δ_i), where $\delta_i = I(T_i \leq C_i)$ indicates whether Y_i is

an uncensored value of T_i, and $Y_i = \min T_i, C_i$.

To model the time-to-event variable T jointly with a longitudinal process, we denote the longitudinal variable that represents that process as Z.

Relationship Between Cox Model and Logistic Regression: In Green & Symon (1983) there is a simple illustration through hazard ratio (HR) and odds ratio (OR) to understand the relationship between the logistic regression model and the Cox model. For simplicity we assume that these models have a single binary covariate X, where $X = 1$ represents a characteristic of interest. For the Cox model, letting T the time to event and $p = \lambda(t \mid X = x) = \lambda_0(t)e^{x\beta}$ then $HR = \frac{p_1}{p_0} = \frac{\lambda(t|X=1)}{\lambda(t|X=0)} = e^{\beta}$. On the other hand, in the logistic regression model, the logit of $\pi = \Pr(X = 1)$ is defined as the natural logarithm of the odds of the survival event of interest: $\ln \frac{\pi}{1-\pi} = X\beta$. The odds ratio is given by $OR = \frac{ODDS|X=1}{ODDS|X=0} = \frac{p/(1-p)|X=1}{p/(1-p)|X=0} = e^{\beta}$. Assuming that the probability of the survival event is very small ($p \to 0$), then the $ODDS = \frac{p}{1-p} \to p$ and therefore:

$$HR = \frac{p_1}{p_0} \approx \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = OR.$$

Since the interpretation of the hazard ratio (HR) is more intuitive than that of the odds ratio (OR), the advantage of the logistic model becomes evident under the analyzed condition. However, the equivalence between Cox and multinomial models, when the event rate is low, is diminished when there are long follow-up times (Green & Symon, 1983).

The conditions for using logistic regression to model time-to-event data also apply to the multinomial model used for competing risks analysis. Although competing risks are characterized by univariate time, they are included in the multivariate survival analysis because there are multiple reasons for follow-up termination (see Houggard, 2000). This adds complexity to the models, further supporting the use of multinomial regression as an alternative for modeling competing risks.

Survival Model in Discrete Time: In order to incorporate the effect of time into the multinomial model, Luo et al. (2016) introduced a spline regression as a covariate. The proposed model has the same structure as the classical multinomial regression model, where the link function is the generalized logit or natural logarithm of the $ODDS$ for the event type δ:

$$\ln\left(\frac{h(t,\delta \mid \vec{x}(t))}{1-h(t \mid \vec{x}(t))}\right) = \Theta\left(t, \vec{x}(t), \vec{\xi}_\delta\right) \tag{1}$$

$$\Theta(t, \vec{x}(t), \vec{\alpha}_\delta, \vec{\beta}_\delta) = \alpha_{1\delta}x_1(t) + \cdots + \alpha_{p\delta}x_p(t) + S(t, \vec{\beta}_\delta),$$

where $\Theta\left(t, \vec{x}(t), \vec{\xi}_\delta\right)$ represents the predictive function, $h(t \mid \vec{x}(t)) = \sum_\delta h(t, \delta \mid \vec{x}(t))$ the overall risk function and $h(t, \delta \mid \vec{x}(t))$ the conditional probability that the δ-type event occurs at time t given covariates and that the event has not yet occurred. Covariates have associated parameter vectors $\vec{\xi}_\delta = \{\vec{\alpha}_\delta, \vec{\beta}_\delta\}$, indicating that each competing risk has its own model. The parameter vector

$\vec{\alpha}_\delta = \{\alpha_{1\delta}, \ldots, \alpha_{p\delta}\}$ is associated with the covariates $\vec{x}(t)$, while β_δ is the parameter vector for a spline regression. The innovation of this proposal lies in the inclusion of a time-dependent cubic spline as a covariate, providing flexibility to the model and capturing irregularities and peaks in the risk accurately. This also partially solves the loss of fit caused by long follow-up times.

$$h(t, \delta \mid \vec{x}(t)) = \frac{\exp\left(\Theta(t, \vec{x}(t), \vec{\xi}_\delta)\right)}{1 + \sum_\delta \exp\left(\Theta(t, \vec{x}(t), \vec{\xi}_\delta)\right)}.$$

For this model, the data must be in an expanded format, where each subject has a record for each observed time point. To estimate the parameters, we assume that there are n independent individuals and consider the following likelihood function, where $\delta = 0, 1, 2$ and $\delta = 0$ represents the reference event. If $k_{it} = \delta_i * I\{t = t_i\}$, where $I\{\cdot\}$ is an indicator function, then:

$$\begin{aligned} L =& \prod_{i=1}^{n} h(t_i, 1 \mid \vec{X}_i(t_i))^{I\{\delta_i=1\}} h(t_i, 2 \mid \vec{X}_i(t_i))^{I\{\delta_i=2\}} \prod_{t=1}^{t_i-1} \left(1 - h(t \mid \vec{X}_i(t_i))\right) \\ =& \prod_{i=1}^{n} \prod_{t=1}^{t_i} \left\{ \frac{\exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1\right)\right)}{1 + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1\right)\right) + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2\right)\right)} \right\}^{I\{k_{it}=1\}} \\ & * \left\{ \frac{\exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2\right)\right)}{1 + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1\right)\right) + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2\right)\right)} \right\}^{I\{k_{it}=2\}} \\ & * \left\{ \frac{1}{1 + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1\right)\right) + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2\right)\right)} \right\}^{I\{k_{it}=0\}}. \end{aligned}$$

3. Survival and Longitudinal Data Joint Models

We have two general models: selection models and mixed models. In the selection model, the joint density function $f_{Z,T}$ is modeled as $f_{T|Z} f_Z$. In the mixed model, it is modeled as $f_{Z|T} f_T$. Both types of models are applicable to longitudinal and survival studies, but mixed models are commonly used for longitudinal studies

with informative dropouts, while selection models are used for survival studies where longitudinal information serves as an event marker. These two approaches can also be combined synergistically. Joint modeling benefits both processes: the survival process helps correct biases in the longitudinal model due to informative dropouts, and appropriate longitudinal adjustments serve as markers for the event. Most studies have focused on models with normal random effects for longitudinal data and proportional hazards models for the time-to-event, both linked with random effects. In this paper, we are particularly interested in the competing risk survival model based on random effects selection models.

The Model and its Maximum Likelihood Estimation: The following example represents a specific case of a joint model. In this case, we have a set of times for subject $i = 1, \ldots, n$, denoted as $\tilde{t}_i = (t_{i1}, \ldots, t_{in_i})$. The longitudinal covariate for subject i is measured at these times $\tilde{t}_i$, and the corresponding values can be represented as $Z_i = (Z_{ij} : t_{ij} \leq T_i)$, where $j = 1, \ldots, n_i$. It's important to be careful and differentiate between the time to event for subject i, denoted as t_i, and the time for the longitudinal response at time j for subject i, denoted as t_{ij}. The covariates in the risk model are assumed to be fixed over time.

$$Z_i(t_{ij}) = b_{0i} + b_{1i}t_{ij} + e_i(t_{ij})$$

$$h_i(t) = h_0(t) \exp\left(\alpha^T \vec{X}_i + \eta(b_{0i} + b_{1i}t)\right).$$

Usually, it is assumed that the errors follow a normal distribution with mean 0 and variance σ_e^2. The random effects $b_i = (b_{0i}, b_{1i})^T$ are also assumed to follow a normal distribution with mean B and variance Σ. As a result, the density function f_Z is a normal distribution with a mean of $b_{0i} + b_{1i}t_{ij}$ and a variance of σ_e^2. Typically, a semiparametric model is used for T.

To simplify the longitudinal model, it is expressed without a fixed component and only considers the random effects b_{0i} and b_{1i}. The inclusion of these random effects in the survival model it is assumes that, when considering the longitudinal history, the risk is influenced by the constant rate of change in the underlying smooth trend (Tsiatis & Davidian, 2004). If the random effects are included as $\eta_1 b_{0i} + \eta_2 b_{1i} + \eta_3(b_{0i} + b_{1i}t)$, the parameters η_1, η_2 and η_3 quantify the association of the longitudinal process with the time to event, induced through the intercept, the slope and the current value of the longitudinal covariate, respectively (Henderson et al., 2000). In the current value effect, a linear trend is assumed for the longitudinal covariate. If $\Omega = \{h_0(\cdot), \alpha, \eta, \sigma_e^2, B, \Sigma\}$ represents the model parameter set, the likelihood for the observed data is:

$$L(\Omega) = \prod_{i=1}^{n} \left[\int_{b_i} \{ \prod_{j=1}^{n_i} f_Z(z_{ij} \mid b_i; \sigma_e^2) \} f_T(Y_i, \delta_i \mid \vec{X}_i, b_i; h_0, \alpha, \eta) f_b(b_i \mid B, \Sigma) \mathrm{d}b_i \right]. \tag{2}$$

In the distribution function f_T, right censoring is assumed according to the following criteria based on the Cox model (see Lawless, 2003):

$$f_T(Y_i, \delta_i \mid \vec{X}_i, b_i; h_0, \alpha, \eta) = \left[h_0(Y_i)\exp(\alpha^T \vec{X}_i + \eta(b_{0i} + b_{1i}t))\right]^{\delta_i} \exp\left[-\int_0^{Y_i} h_0(u)\exp(\alpha^T \vec{X}_i + \eta(b_{0i} + b_{1i}t))\mathrm{d}u\right]. \quad (3)$$

According to Tsiatis & Davidian (2004), the main assumptions the likelihood has in (2) are: a) conditional on the random effects, Z and T are independent, b) the censoring mechanism and the longitudinal calendar are noninformative, c) The longitudinal model does not consider the autocorrelation structure and focuses on the relationship between the smooth trend and the time to event, d) random effects are independent of the covariates $\vec{X}$, and e) the errors are mutually independent and independent of all other variables.

For the estimation of semiparametric joint models, Wulfsohn & Tsiatis (1997) introduced a method that maximizes the likelihood using an EM algorithm. This algorithm has been widely adopted in recent proposals for the estimation process.

A recommended approach for joint estimation involves starting with the well-known two-stage model, where the longitudinal and survival models are estimated separately. The random effects obtained from the longitudinal model are then used to estimate the survival model. While the two-stage model can result in biased estimates for the joint model, the estimates obtained at this stage serve as a useful starting point for maximizing the likelihood of the joint model. This maximization can be achieved using EM algorithms, such as the one proposed by Wulfsohn & Tsiatis (1997).

Properties of the Estimators: Zeng & Cai (2005) provided rigorous proof, assuming normality of the random effects, of the strong consistency of maximum likelihood estimators for joint models of repeated measurements and survival time. They also derived the asymptotic distributions of these estimators, which follow a multivariate normal distribution. The asymptotic results remain valid even if the random effects exhibit slightly heavier tails than the normal density. Furthermore, maximizing the joint likelihood leads to unbiased estimates for joint models.

The likelihood maximization procedures for parameter estimation and property description are typically performed for semiparametric joint models, where the survival model is semiparametric. However, these results can also be extended to fully parametric joint models.

4. Joint Model of Competitive Risks in Discrete Time with Longitudinal Information

This paper proposes an extension to the model introduced by Luo et al. (2016) using a joint modeling approach. The data structure is illustrated in Table 1, which shows three different event cases with the corresponding time-to-event, a baseline covariate X, and a longitudinal variable Z. The data structure is similar to the expanded framework proposed by Luo et al. (2016), but it is utilized differently in our model. In our likelihood equations (6)-(9), the longitudinal information

is not included in the multinomial model for each data line in the longitudinal calendar. Instead, it is summarized using random effects, resulting in a single line of information per individual in the survival model.

The extension is easily achieved by replacing the Cox model with the multinomial model in the joint model. We propose a dependency on random effects for the longitudinal variable while ignoring fixed effects. When the goal is survival analysis, including the fixed part is not necessary. The specific choice of a quadratic trend for Z is made to illustrate the model in the next section.

$$Z_i(t_{ij}) = b_{0i} + b_{1i}t_{ij} + b_{2i}t_{ij}^2 + e_i(t_{ij}) \tag{4}$$

$$\ln\left(\frac{h_i(t,\delta \mid \vec{x}_i, b_i)}{1 - h_i(t \mid \vec{x}_i, b_i)}\right) = \Theta\left(t, \vec{x}_i, b_i, \vec{\alpha}_\delta, \vec{\eta}_\delta, \vec{\beta}_\delta\right) \tag{5}$$

$$\Theta(t, \vec{x}_i, b_i, \vec{\alpha}_\delta, \vec{\eta}_\delta, \vec{\beta}_\delta) = \alpha_{1\delta}x_{1i} + \cdots + \alpha_{p\delta}x_{pi} + B(b_i, \vec{\eta}_\delta) + S(t, \vec{\beta}_\delta).$$

Table 1: Format of the joint model database

Case	Time	Event	Covariate	Longitudinal
1	1	1	X_1	Z_{11}
1	2	1	X_1	Z_{12}
1	3	1	X_1	Z_{13}
1	4	1	X_1	Z_{14}
2	1	2	X_2	Z_{21}
2	2	2	X_2	Z_{22}
2	3	2	X_2	Z_{23}
3	1	0	X_3	Z_{31}
3	2	0	X_3	Z_{32}
3	3	0	X_3	Z_{33}
3	4	0	X_3	Z_{34}
3	5	0	X_3	Z_{35}

The term $B(b_i, \vec{\eta}_\delta)$ captures the various combinations that the effect of the longitudinal variable can have on the time to event through its random effects. Here, all components of $\vec{X}$ are baseline covariates and may differ for each competing risk category.

In equation (6) for $L(\Omega)$, both f_Z and f_b maintain the same normality structure as described in the previous section. However, f_T now depends on the spline parameters and undergoes a significant change. Specifically, takes the form of a multinomial distribution, which is illustrated in a particular case where $\delta = 0, 1, 2$.

$$L(\Omega) = \prod_{i=1}^{n}\left[\int_{b_i}\{\prod_{j=1}^{n_i} f_Z(z_{ij} \mid b_i; \sigma_e^2)\} f_T(Y_i, \delta_i \mid \vec{X}_i, b_i; h_0, \alpha, \eta, \beta) f_b(b_i \mid B, \Sigma)\mathrm{d}b_i\right] \tag{6}$$

$$f_b(b_i \mid B, \Sigma) = (2\pi|\Sigma|)^{-1/2}\exp\{-0,5(b_i - B)^T\Sigma^{-1}(b_i - B)\} \tag{7}$$

$$f_Z(z_{ij} \mid b_i; \sigma_e^2) = (2\pi\sigma_e^2)^{-1/2} \exp\{-0,5(z_i(t_{ij}) - b_{0i} - b_{1i}t_{ij} - b_{2i}t_{ij}^2)^2/\sigma_e^2\} \quad (8)$$

$$f_T(Y_i, \delta_i \mid \Omega_T) = \left\{ \frac{\exp\left(\Theta\left(Y_i, \vec{x}_i(\tilde{t}_i), b_i, \vec{\alpha}_1, \vec{\eta}_1, \vec{\beta}_1\right)\right)}{1 + \exp\left(\Theta\left(Y_i, \vec{x}_i(\tilde{t}_i), b_i, \vec{\alpha}_1, \vec{\eta}_1, \vec{\beta}_1\right)\right) + \exp\left(\Theta\left(Y_i, \vec{x}_i(\tilde{t}_i), b_i, \vec{\alpha}_2, \vec{\eta}_2, \vec{\beta}_2\right)\right)} \right\}^{I\{\delta_i=1\}}$$
$$* \left\{ \frac{\exp\left(\Theta\left(t_i, \vec{x}_i(t_i), b_i, \vec{\alpha}_2, \vec{\eta}_2, \vec{\beta}_2\right)\right)}{1 + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), b_i, \vec{\alpha}_1, \vec{\eta}_1, \vec{\beta}_1\right)\right) + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), b_i, \vec{\alpha}_2, \vec{\eta}_2, \vec{\beta}_2\right)\right)} \right\}^{I\{\delta_i=2\}} \quad (9)$$
$$* \left\{ \frac{1}{1 + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), b_i, \vec{\alpha}_1, \vec{\eta}_1, \vec{\beta}_1\right)\right) + \exp\left(\Theta\left(t_i, \vec{x}_i(t_i), b_i, \vec{\alpha}_2, \vec{\eta}_2, \vec{\beta}_2\right)\right)} \right\}^{I\{\delta_i=0\}}$$

where $I\{\cdot\}$ is an indicator function. Although equations (3) and (9) are different, they share a similar essence. In (3), the contribution of each individual is included in the Cox model based on whether they have observed the event or not. In (9), the contribution of each individual is included based on the odds of the competing risk to which they belong.

There are various proposals for fitting joint models using standard software, but the R packages FastJM (Li et al., 2022), JM, and JMbayes2 (Rizopoulos, 2010, 2016), are the ones that allow modeling longitudinal and survival data with competing risks. However, these packages utilize a semiparametric model for competing risk and cannot be used to fit our fully parametric joint model. These packages are continuously evolving and expanding their capabilities for joint modeling. It is only a matter of time before other proposals, including ours, are incorporated into these packages. In the meantime, when fitting models that are not directly supported by standard software, approaches such as the one proposed by Guo & Carlin (2004) can be utilized. One method from their proposal involves adapting the SAS NLMIXED procedure (SAS Institute, 2013) to fit joint models. This adaptation is straightforward and can only be used with fully parametric survival models. In our case, we utilize this approach by adding the log-likelihood of a multinomial model to the log-likelihood of the longitudinal model with normal errors and normal random effects.

Joint models pose computational challenges as they involve estimating two processes simultaneously. While NLMIXED is a powerful tool for optimizing likelihoods and estimating nonlinear mixed models, it may not provide immediate results for datasets with a large number of individuals, such as our application with 912 individuals (on Core i7 computers with 12 MB of RAM). Furthermore, the complexity of the likelihood function can result in different outcomes depending on the starting point used for numerical maximization. Estimation may require an iterative process to find the optimal maximization point, which can be time-consuming. To mitigate computation time, we initially use the two-stage model mentioned earlier and then maximize the joint likelihood based on the outcomes obtained from it.

5. Application

Although our model can be applied to any type of data, we will illustrate it using a similar example to the one used by Luo et al. (2016). The dataset consists of 912 individuals who took out a 36-month consumer loan. Each individual can terminate the credit for three reasons: maturity ($\delta = 0$), default ($\delta = 1$), or debt prepayment ($\delta = 2$). The first and third events represent the attrition risk, meaning that the customer pays off their balance. If the customer does not fall into any of these three categories, their credit remains open. To simplify the model, these individuals are included in the maturity group, assuming that they are all regular customers. The dataset includes variables such as the duration of the credit and the reason for its termination, baseline covariates like sex and age at the beginning of the credit, and a longitudinal variable representing the monthly balance of the customer's main credit card.

The example involves a competing risks scenario, where the end of the loan represents the time to event, which can occur due to three different causes. Although we can determine the probabilities for all three events, our analysis focuses on the default event. We model the time to loan termination using multinomial logistic regression along with a longitudinal variable. In joint models for survival analysis, a longitudinal variable is typically used as a marker for the event of interest. However, in this discrete time problem, we include the longitudinal variable to appropriately estimate the probability of default, assuming that it captures the potential variations in risk. To ensure smooth trends in these probabilities, we incorporate smoothed longitudinal trajectories into the model using kernel regression techniques (refer to Bowman & Azzalini, 1997).

Before presenting the results of our proposed model, let's first discuss the results obtained from the Cox model and the model by Luo et al. (2016). In the Cox model, we treated prepayment and maturity times as censored data. Although it is not appropriate to ignore competing risks and the longitudinal variable, we did this to illustrate the outcomes of the most basic model. We included age as a covariate, and the baseline risk was estimated using the kernel estimator (refer to Klein & Moeschberger, 2005). The nonparametric nature of the baseline risk estimation allowed us to overlook the truncation effect in the data. From the

estimation results shown in Figure 1, three important findings can be highlighted. First, there is a typical hump-shaped trend in the default risk, indicating that the risk initially increases, reaches a peak, and then decreases over time. Second, there is an inverse relationship between age and default risk, meaning that younger individuals have a higher risk of default compared to older individuals. Finally, the estimated default risk aligns with the overall kernel estimation at the average age of 32 years.

On the other hand, when using a logistic model with age as the only covariate, we obtained an estimated parameter of 0.036, which is quite close to the value of 0.032 obtained in the Cox model. The similarity in the estimates and the hazard ratio (HR) and odds ratio (OR) derived from them suggest that the logistic model is a suitable alternative for modeling the time to default.

According to the estimated Cox model, there are slight peaks at times 6 and 16 (see Figure 1). However, when comparing an adjusted model that considers these peaks with another model that only includes the effect of the spline associated with a single parameter β, the estimated probabilities are practically equal. To keep the model simple, we include the time effect only with the spline, which can be interpreted as the effect of elapsed time since the start of the credit. The predictive function for the model by Luo et al. (2016), defined in Equation (1), with only the significant covariates is as follows:

$$\Theta(t, \vec{x}(t), \vec{\alpha}_\delta, \beta_\delta) = \alpha_{0\delta} + \alpha_{1\delta} Age(t) + \alpha_{2\delta} CreditCard(t) + \beta_\delta S(t).$$

In this model, similar to the joint model, $S(t)$ represents a cubic spline of the event time.

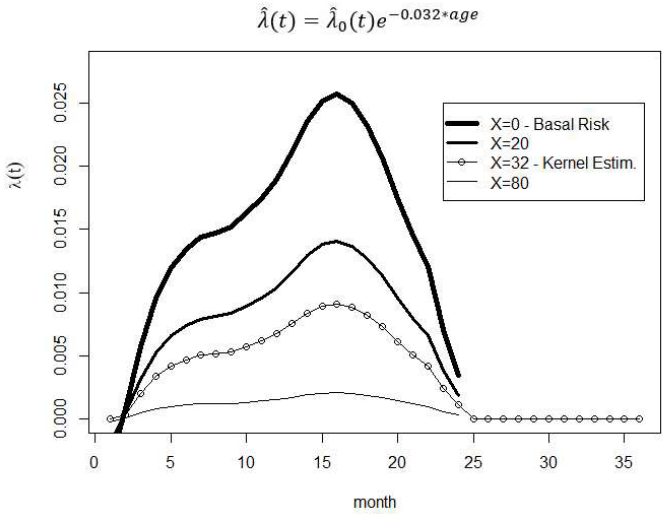

FIGURE 1: Cox regression model for default time

The parameter estimation for the Luo et al. (2016) model, focusing solely on the time to default, is presented in Table 2. It is worth noting that the estimation

of the parameter associated with age differs when considering multiple lines of information per individual, as opposed to the joint model where the age effect is measured only once per individual (Table 3).

TABLE 2: Luo et al. (2016) model for the time to default

Parameter	Estimation	Standard Error	p-value
Intercept α_0	11,0382	1,1754	$<$ 0,0001
Age α_1	-0,0306	0,0134	0,0226
Credit Card α_2	-356,450	30,742	$<$ 0,001
Spline β	-0,0784	0,0130	$<$ 0,001

The first part of Figure 2 depicts the credit card balances for three individuals aged 32, while the second part illustrates the corresponding default risk. We select these people of that age because they are in the average risk trend to default. We observe an inverse relationship between the balance and the risk, where higher balances are associated with a lower risk of default. The most notable case is individual 2, who has been on credit for 27 months and experiences a zero estimated risk starting from the ninth month. Individual 3 exhibits a contradictory result, with the risk decreasing almost to zero and then increasing during periods without any default cases. Therefore, we can conclude that the estimated model has difficulties in accurately determining the default risk.

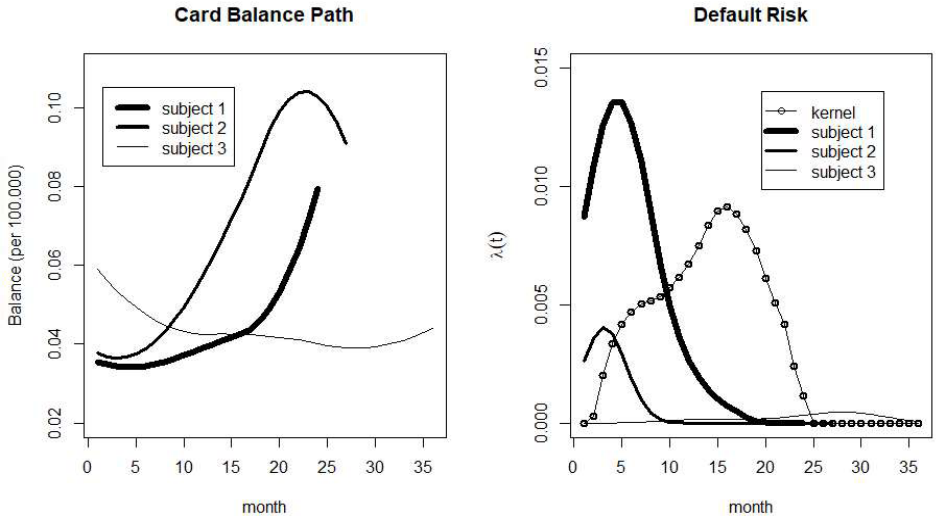

FIGURE 2: Longitudinal variable trajectories and default risk under Luo et al. (2016) model for 3 subjects

The inverse relationship between debt and default risk is intriguing as it appears to go against common intuition. According to the obtained relationship, even young individuals with high debt have low default risks, which seems contradictory. However, this can be attributed to the fact that the high balances reflect a high purchasing power of the customers included in the database. This aspect partially explains the unexpected result.

The joint model applied to this dataset corresponds to the one mentioned in equations (4)-(5). The significant variables in the proposed survival model include baseline age, random effects of the credit card balance incorporated as the current value, and a regression spline to adjust for irregular risk components (refer to Table 3). Therefore, the predictive function of the model can be defined as follows:

$$\Theta(t, x_i, b_i, \vec{\alpha}_\delta, \eta_\delta, \beta_\delta) = \alpha_{0\delta} + \alpha_{1\delta} Age_i + \eta_\delta(b_{0i} + b_{1i}t_{ij} + b_{2i}t_{ij}^2) + \beta_\delta S_i(t_{ij}).$$

TABLE 3: Joint model of the evolution of credit card balance and default time.

Parameter	Estimation	Standard Error	p-value
Longitudinal			
B_0	0,057199	0,000710	$<$ 0,0001
B_1	-0,000071	0,000032	0,0278
B_2	-0,000002	0,000001	0,0455
σ_{11}	0,023510	0,000056	$<$ 0,0001
σ_{12}	-0,000624	0.000051	$<$ 0,0001
σ_{13}	0,000052	0.000620	0,9332
σ_{22}	0,003073	0.000078	$<$ 0,0001
σ_{23}	-0,000132	0.004766	0,9779
σ_{33}	0,000006	0.000002	0,0069
σ_e^2	0,000403	0.000168	0,0164
Default T_1			
Intercept α_{10}	-1,000	0,52611	0,0574
Age α_{11}	-0,026	0,00458	$<$ 0,0001
Current Value η_1	-48,0	13,3333	0,0003
Spline β_1	-0,059	0,01283	$<$ 0,0001
Prepayment T_2			
Intercept α_{20}	-4,998	2,37619	0,0357
Age α_{21}	0,040	0,01250	0,0015
Current Value η_2	-0,372	0,10625	0,0005
Spline β_2	0,045	0,01030	$<$ 0,0001

Based on the estimated joint model presented in Table 3, we observe an inverse relationship between the covariates and default risk. Although the individual trajectories of the card balance are not shown in the table due to space constraints, they were estimated using the empirical Bayesian estimator of the random effects (refer to Verbeke & Molenberghs 2000, section 7.2). With these trajectories, we can estimate default probabilities, which are illustrated in Figure 3 using a similar approach as in Figure 2 and featuring the same individuals. Figure 3 provides insights into how the longitudinal variable influences the default risk trend and even modifies its overall hump shape in certain cases.

Although we have observed different individual estimates for the default risk in both models, they have a similar level of goodness of fit. The Nagelkerke pseudo R^2 (Nagelkerke, 1991) for the model by Luo et al. (2016) is 0.298, while for the joint model it is 0.322. This suggests that the longitudinal model is accurately capturing the underlying trend. Any loss of fit caused by the variability that the model cannot capture is compensated by including the baseline variable appropriately, resulting in an overall satisfactory fit.

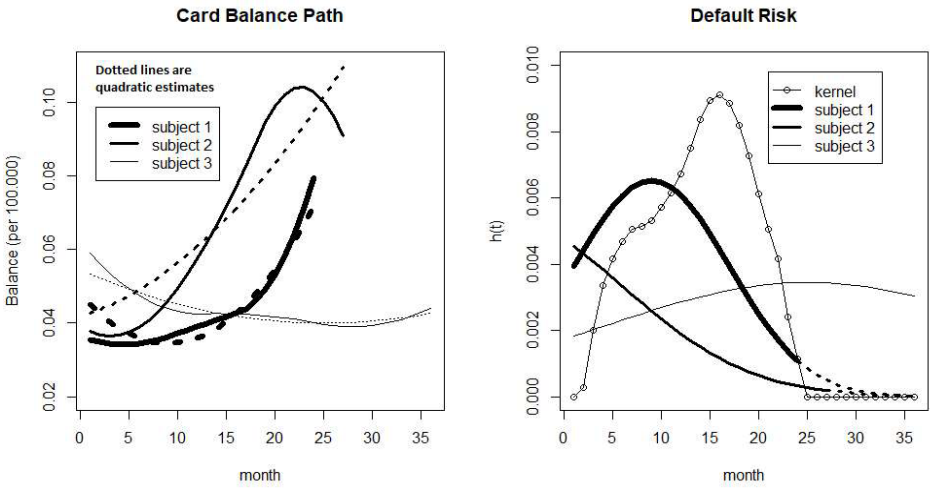

FIGURE 3: Longitudinal variable trajectories and risk of default under a joint model for three individuals

In conclusion, the covariates have an impact on the default risk as follows: The trend is determined by the longitudinal variable, which increases or decreases proportionally over time depending on the baseline age and duration of the debt. Generally, the projections obtained from the joint model are closer to the overall trend of the risk determined by the kernel estimator compared to the results from the model proposed by Luo et al. (2016). However, both models have quite similar goodness of fit and limitations. They can generate probabilities greater than zero at the end of the credit period where no default cases were observed, and the probabilities can be high if the balance of the card is very low, although such cases are rare.

Default probabilities play a crucial role for finance companies in establishing reserves to address potential default issues among their clients. These probabilities are based on the time elapsed from the start of the debt to the maturity date. It's important to note that individuals whose credits have already reached maturity are included in the model solely for estimation purposes, and no reserves are set aside for them. When using the classic Cox and logistic models to assign these probabilities, each person is assigned a unique probability based on the average trend of their covariates. However, alternative models such as our joint model proposal or the one proposed by Luo et al. (2016) can provide different probabilities for each individual based on their projected outcomes. This allows for a more tailored and accurate assessment of default risk for each client.

6. Discussion

When modeling competing risks with longitudinal information, the model proposed by Elashoff et al. (2007) and similar ones are typically the first alternatives

to consider. However, these models have some disadvantages when it comes to discrete time modeling, and they can be complex and not readily available in standard software. On the other hand, if the conditions are met to apply the logistic model to the problem, the model proposed by Luo et al. (2016) can be used with standard software. However, it's important to note that simultaneous inclusion of both baseline and longitudinal covariates is not suitable under this model when there are customers with ongoing credits. This can create an imbalance in the management of baseline covariates, affecting the accuracy and interpretation of the results.

Joint models for survival data with longitudinal information offer a solution for modeling competing risks by incorporating both longitudinal and baseline information. However, one drawback of these models is the complexity of the likelihood function, which can complicate the estimation process. If additional longitudinal variables are included as covariates to explain the risk, it will further increase the computational demand of the optimization process. Moreover, the inclusion of multiple longitudinal variables requires considering complex interactions between them, which are not discussed in this paper but can be found in other references such as Liu & Huang (2009) and Lin et al. (2022).

[Received: February 2023 — Accepted: June 2023]

References

Annesi, I., Moreau, T. & Lellouch, J. (1989), 'Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies', *Statistics in Medicine* **8**, 1515–1521.

Begg, C. B. & Gray, R. (1984), 'Calculation of polychotomous logistic regression parameters using individualized regressions', *Biometrika* **71**(1), 11–18.

Bowman, A. W. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations*, 1 edn, New York, Oxford University Press Inc.

Cox, D. R. (1972), 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Elandt-Johnson, R. C. (1980), Time dependent logistic models in follow-up studies and clinical trials, I. Binary data, Technical report, Institute of Statistics Mimeo Series No 1309, University of North Carolina.

Elashoff, R. M., Li, G. & Li, N. (2007), 'An approach to joint analysis of longitudinal measurements and competing risks failure time data', *Statistics in Medicine* **26**, 2813–2835.

Elashoff, R. M., Li, G. & Li, N. (2008), 'Joint model for longitudinal measurements and survival data in the presence of multiple failure types', *Biometrics* **64**, 762–771.

Green, M. S. & Symon, M. J. (1983), 'A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies', *Journal of Clinical Epidemiology* **36**, 715–723.

Guo, X. & Carlin, B. P. (2004), 'Separate and joint modeling of longitudinal and event time data using standard computer packages', *The American Statistician* **58**, 16–24.

Henderson, R., Diggle, P. & Dobson, A. (2000), 'Joint modelling of longitudinal measurements and event time data', *Biostatistics* **4**, 465–480.

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

Klein, J. P. & Moeschberger, M. L. (2005), *Survival Analysis: Techniques for Censored and Truncated Data*, 3rd edn, Springer, New York.

Lawless, J. F. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd edn, Wiley, Hoboken.

Li, N., Elashoff, R. & Li, G. (2009), 'Robust joint modeling of longitu-dinal measurements and competing risks failure time data', *Biometrical Journal* **51**, 19–30.

Li, S., Li, N., Wang, H., Zhou, J., Zhou, H. & Li, G. (2022), 'Efficient algorithms and implementation of a semiparametric joint model for longitudinal and competing risk data: With applications to massive biobank data', *Computational and Mathematical Methods in Medicine* **2022**, 1–12.

Lin, H., McCulloch, C. E. & Mayne, S. (2022), 'Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables', *Statistics in Medicine* **21**(16).

Liu, L. & Huang, X. (2009), 'Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome', *Journal of the Royal Statistical Society* **58**, 65–81.

Luo, S., Kong, X. & Nie, T. (2016), 'Spline based survival model credit risk modeling', *European Journal of Operational Research* **253**, 869–879.

Myers, M. H., Hankey, B. F. & Mantel, N. A. (1973), 'A logistic-exponential model for use with response time data involving regession variables', *Biometrics* **29**, 257–296.

Nagelkerke, N. (1991), 'A note on a general definition of the coefficient of determination', *Biometrika* **78**, 691–692.

Rizopoulos, D. (2010), 'JM: An R package for the joint modelling of longitudinal and time-to-event data', *Journal of Statistical Software* **35**, 1–33.

Rizopoulos, D. (2016), 'The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC', *Journal of Statistical Software* **72**(7), 1–45.

Salazar, A. M. (2021), Un modelo conjunto de riesgos competitivos en tiempo discreto con información longitudinal, Tesis de maestría, Universidad Nacional de Colombia, Facultad de Ciencias. Departamento de Estadística, Bogotá.

SAS Institute (2013), 'Sas/stat 9.4'. Cary, North Carolina, United States, SAS Institute Inc.

Teixeira, L., Sousa, I., Rodriguez, A. & Mendonça, D. (2019), 'Joint modelling of longitudinal and competing risks data in clinical research', *REVSTAT-Statistical Journal* **17**(2), 245–264.

Tsiatis, A. A. & Davidian, M. (2004), 'Joint modeling of longitudinal and timeto-event data: An overview', *Statistica Sinica* **14**, 809–834.

Verbeke, G. & Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, 1 edn, Springer-Verlag, New York.

Williamson, P. R., Kolamunnage-Dona, R., Philipson, P. & Marson, A. G. (2007), 'Joint modelling of longitudinal and competing risks data', *Statistics in Medicine* **27**, 6426–6438.

Wulfsohn, M. & Tsiatis, A. (1997), 'A joint model for survival and longitudinal data measured with error', *Biometrics* **53**, 330–339.

Zeng, D. & Cai, J. (2005), 'Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time', *The Annals of Statistics* **33**, 2132–2163.