Revista Brasileira de Computação Aplicada, November, 2024

DOI: 10.5335/rbca.v16i3.15375
Vol. 16, Nº 3, pp. 1–9
Homepage: seer.upf.br/index.php/rbca/index

UPF EDITORA
RBCA ISSN 2176-6649

ORIGINAL PAPER

Simulation study of viable scenarios for potential numerical convergence issues in fitting joint models for longitudinal and survival data

Daiane de Oliveira Gonçalves [1], Natália da Silva Martins Fonseca [2], Marcelo Ângelo Cirillo [3]

[1]Ph.D. student in Statistics and Agricultural Experimentation, Federal University of Lavras, [2]Department of Statistics, Federal University of Alfenas, [3]Department of Statistics, Federal University of Lavras

*prof.daiane.oliveira@gmail.com; natalia.martins@unifal-mg.edu.br; macufla@ufla.br

Received: 2023-11-16. **Revised:** 2024-07-12. **Accepted:** 2024-10-19.

Abstract

Studies concerning the characteristics of phenomena/experiments over time, such as longitudinal studies or those focused on the time until an event of interest occurs, are increasingly essential in various fields. There may be instances where the investigation of the relationship between one or more longitudinal responses and an event of interest is warranted, a task achievable through the joint model of longitudinal and survival data. However, these models may have convergence problems and be computationally demanding, making their use unfeasible in many cases. In consideration of these factors, the objective of this study is to conduct a Monte Carlo simulation study involving various censoring percentages and correlation structures. The proposed cross-coverage probability will be employed as a diagnostic tool to identify circumstances conducive to numerical convergence, aiming to obtain maximum likelihood estimates for joint models applied to longitudinal and survival data. The results indicated similarity in terms of inference among the models, accounting for the impact of both the correlation structure and the censoring percentage. It was determined that the cross-coverage probability contributes to diagnosing the favorable behavior of the data, thereby facilitating the implementation of joint modeling.

Keywords: Censorship; longitudinal data; mixed linear models; simulation; survival analysis

Resumo

Estudos relacionados a características de fenômenos/experimentos no tempo, como estudos longitudinais ou do tempo até a ocorrência de um evento de interesse, se fazem cada vez mais necessários em diversas áreas. Podem existir situações em que se objetiva investigar a relação entre uma ou mais respostas longitudinais e um evento de interesse, que pode ser realizada com o auxilio da modelagem conjunta de dados longitudinais e de sobrevivência. Entretanto, esses modelos podem apresentar problemas de convergência e serem computacionalmente exigentes, tornando inviável a utilização dos mesmos em muitos casos. Tendo em vista esses fatores, o objetivo deste trabalho é realizar um estudo de simulação de Monte Carlo envolvendo diversas percentagens de censura e estruturas de correlação. A probabilidade de cobertura cruzada proposta será utilizada como ferramenta de diagnóstico para identificar circunstâncias favoráveis à convergência numérica, visando à obtenção de estimativas de máxima verossimilhança para modelos conjuntos aplicados a dados longitudinais e de sobrevivência. Como resultados, verificou-se a existência similaridade em termos de inferência entre os modelos, com efeito da estrutura de correlação e do percentual de censura. Constatou-se que a probabilidade de cobertura cruzada contribui com um diagnóstico sobre o bom comportamento dos dados, auxiliando para realização da modelagem conjunta.

Palavras-Chave: Censura; dados longitudinais; modelos lineares de efeitos mistos; simulação; análise de sobrevivência

1 Introduction

Many pieces of information are currently collected over time, known as longitudinal data, obtained from the same sample elements over an extended period. Longitudinal data represent repeated observations of a random variable of interest, collected at different time points for the same individual or object (Hu and Szymczak, 2023).

In statistics, numerous methodologies are available for analyzing such data. Among these techniques, mixed-effects linear and survival models stand out, with the latter being particularly useful when dealing with censored data (incomplete observations of the response variable).

Mixed-effects linear models are defined as models that include both fixed effects and random effects. They are primarily used to describe the relationship between a response variable and covariates in data grouped according to one or more classification factors (Pinheiro and Bates, 2006).

These models enable the prediction of how individual response trajectories change over time and the estimation of parameters describing how the mean response changes in the population of interest. They can accommodate any degree of imbalance in the data, meaning that the number of measurements does not need to be the same for each individual or object. Additionally, random effects account for the correlation between repeated measures in a relatively efficient manner (Verbeke et al., 1997).

Survival models are designed for situations where the goal is to evaluate the time until the occurrence of one or more events of interest, often referred to as failures. However, the exact time of occurrence of the event of interest is not always known, or the event may not be observed at all, leading to censoring in survival models. Censored observations are partial or incomplete observations of the response variable (Colosimo and Giolo, 2006).

Thus, survival models are distinguished by their capacity to accommodate these incomplete (censored) observations in analysis, thereby enabling robust statistical conclusions by incorporating information about the time until the occurrence of the event of interest for the sampled elements.

There is also the possibility to investigate the relationship between one or more longitudinal responses and an event of interest. The statistical treatment of responses repeated over a period of time and observed in the same experimental unit can be applied in different situations involving specific models. In view of the above and given a longitudinal study considering n individuals, the use of a joint model (Viviani et al., 2014) allows the time until the occurrence of an event of interest to be modeled, including covariates that vary over time. In this case, Wu and Carroll (1988) suggest joint modeling using survival analysis techniques with random effects models.

The relationship between the mixed linear models with analysis of survival data such that random effects act linearly on the survival time of the individual or experimental unit is mentioned by Do Ha and Lee (2005). Rizopoulos (2012) includes random effects in survival data, allowing for the prediction of dynamic individual response trajectories over the observed period.

A joint model that simultaneously contemplates the longitudinal responses in the presence of censoring has been proposed. Zhang et al. (2014) recommend applying this in situations represented by survival models with measurement errors, missing data with time-dependent covariates and longitudinal models. However, in many cases, the numerical complexity of fitting these models can make them unfeasible since including random effects becomes computationally demanding as their dimensionality increases (Murray and Philipson, 2022).

Notably, the longitudinal process and the survival process are associated with latent variables. In this context, Rizopoulos and Lesaffre (2014) highlight that models with latent variables are defined based on the assumption of conditional independence. In practice, these models are difficult to implement since the specified integral with respect to the latent variable does not have a form. Therefore, numerical integration is needed, making these models very computationally demanding.

Another important issue is mentioned by Rizopoulos (2010): considering the accelerated time to failure, the specification of the joint model requires a complete longitudinal history for calculating the survival function and the risk function; in many applications, individuals and/or units may exhibit highly nonlinear longitudinal trajectories.

Given the previous description and considering the convergence problems that may occur, the use of latent variables and their implications in solving the integral with required computational demand, preliminarily evaluating the behavior of the data through individual process modeling of survival and longitudinal is worth investigating since similar parameter estimation results may otherwise occur. Therefore, the performance of a joint model can be better analyzed than that of other models.

This perspective justifies the contributions of this study, which presents a methodology that obtains the cross coverage probability. In the proposed methodology, the estimates of the longitudinal model parameters are computed based on the confidence interval of the parameters of the survival model. Thus, the coverage probabilities for the survival model are generated by inverting the intervals.

The main contribution of this work is the introduction and application of cross-coverage probability as a diagnostic tool. This tool is employed to identify circumstances conducive to numerical convergence in obtaining maximum likelihood estimates for joint models applied to longitudinal and survival data. This diagnostic significantly aids in overcoming convergence issues and the computational demands often associated with these models, thereby enhancing applicability in studies involving such data types and yielding more precise results while leveraging the advantages these models offer.

Notably, the coverage probabilities in both models were not computed based on the parametric values. Rather, they were computed considering sample estimates, intuitively the bootstrap approach, in which the sample estimate is considered a parametric value for interval confidence levels estimated in the subsamples.

In view of the above, this study proposes using the measure of the probability of cross-coverage as a diagnostic tool for connecting longitudinal and survival models. This can help the researcher estimate a joint model that involves both processes and minimize possible numerical convergence problems.

2 Materials and methods

For a better compression of the construction of the panel of data with repeated measures in the absence and presence of censure, as well as the notation used in the subsequent sections, the layout described in Table 1 is followed.

Table 1: Panel data layout with repeated measures ($m = 1, \cdots, M$), within each group ($g = 1, \cdots, G$) censored (δ).

Longitudinal Process			Survival Process	
Y	G	X	W	δ
y_{11}	1	x_{11}	w_{11}	δ_{11}
y_{21}	1	x_{21}	w_{21}	δ_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{m1}	1	x_{m1}	w_{m1}	δ_{m1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{1g}	g	x_{1g}	w_{1g}	δ_{1g}
y_{2g}	g	x_{2g}	w_{2g}	δ_{2g}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{mg}	g	x_{mg}	w_{mg}	δ_{mg}

The longitudinal process and simulated survival, including the categorical covariates based on this structure are described below in sections 2.1 – Monte Carlo simulation of the multilevel model for the longitudinal process; 2.2 – Monte Carlo simulation of the Weibull model for the survival process; 2.3 – Definition of the simulation scenarios and parametric values; and 2.4 – Adjustment of the models for the survival and longitudinal processes with inclusion of categorical covariates and estimation of the probabilities of cross coverage.

2.1 Monte Carlo simulation of the multilevel model for the longitudinal process

Y was assumed to be the dependent variable in the fit of the multilevel model with the distribution $Y_j \sim N_p\left(\mu_{jg}, \Sigma_a\right)$, for $j = 1, \cdots, M \cdot G$, where $g = 1, \cdots, G$ such that the dependence relationship with the regressor variable X was maintained by the Eq. (1) (Silva and Cirillo, 2018) and

$$\mu_{jg} = \beta_0\left(m_j - 1\right) + \beta_1 X_{jg}, \qquad (1)$$

in which $X_{jg} \sim U(0,1) and \beta_0 = \beta_1 = 0.5$, fixed arbitrarily.

The autoregressive correlation structure of order 1, AR(1), was considered for the definition of the covariance matrix Σ_a, where $a = 1$. Its estimated correlations were given as a function of the α parameters used in the generalized estimation equations approach (Liang and Zeger, 1986) (2)

$$CORR\left(Y_{(g,j)}, Y_{(g,j+t)}\right) = \alpha^t, \text{where} \quad t = 1, \cdots, T. \qquad (2)$$

For $\sum_2(a = 2)$, we proceeded by including the interchangeable correlation structure, according to the Eq. (3).

$$CORR\left(Y_{gj}, Y_{gj'}\right) = \begin{cases} 1, & \text{if } j = j' \\ \alpha, & \text{if } j \neq j' \end{cases}. \qquad (3)$$

The inclusion of the degree of correlation ρ in the estimates of α in Eq. (2) and Eq. (3) was performed using the method for obtaining the limiting estimates of the covariance matrix proposed by Silva and Cirillo (2018). This method was applied to the GEE 2 models according to Eq. (4) and Eq. (5).

$$\alpha_0(1-\alpha_0)^{-1}\left\{\frac{t-(1-\alpha_0^t)}{1-\alpha_0}\right\} - t(t-1)\frac{\rho}{2} = 0, \qquad (4)$$

where $-1/(t-1) \leq \rho \leq 1$, and

$$\alpha_0 = 2\rho\left\{\frac{t-(1-\rho^t)/(1-\rho)}{t(t-1)(1-\rho)}\right\}, \qquad (5)$$

where $-1 \leq \rho \leq 1$.

The restriction presented in Eq. (4) is performed assuming that the exchangeable correlation matrix is true when considering it as a working correlation matrix; analogously, applying the restriction presented in Eq. (5) assumes the AR(1) structure to be true (Sutradhar and Das, 2000).

2.2 Monte Carlo simulation of the survival model

The survival process was simulated so that the percentage of censorship was controlled in the generated sample. To do so, the procedure used in Giarola et al. (2018) assumed two auxiliary random variables, $W_1 \sim Weibull(\alpha_1, \beta_1)$ and $W_2 \sim Weibull(\alpha_2, \beta_2)$. In this way $Z = W_2 - W_1$ was defined with the condition that $\alpha_1 = \alpha_2 = \alpha_c$. Substituting this into F_Z, we obtained in Eq. (6)

$$\begin{aligned} F(z) &= \int_0^\infty w_1^{\alpha_c - 1} - \exp\left\{-\left(\frac{w_1}{\beta_1}\right)^{\alpha_c} - \left(\frac{w_2}{\beta_2}\right)^{\alpha_c}\right\} dw_1 \\ &= \frac{1}{\alpha_c\left(\frac{1}{\beta_1^{\alpha_c}} + \frac{1}{\beta_2^{\alpha_c}}\right)}. \end{aligned} \qquad (6)$$

Therefore,

$$F_z = \frac{\beta_1^{\alpha_c}}{\beta_1^{\alpha_c} + \beta_2^{\alpha_c}}. \quad (7)$$

Thus, given that W_2 considered the censoring time associated with the i-th observation and W_1 considered the failure time, the definition of the censoring percentage P was determined by Eq. (8)

$$P = \frac{\beta_{1g}^{\alpha_c}}{\beta_{1g}^{\alpha_c} + \beta_{2g}^{\alpha_c}}, \text{where} \quad g = 1, \cdots, G, \quad (8)$$

$$\beta_{2g}^{*} = \beta_{1g}\left(\frac{1-P}{P}\right)^{\frac{1}{\alpha_c}}. \quad (9)$$

Following these specifications, the censoring assignment was given by generating $F \sim Weibull(\alpha_g, \beta_g)$, where $g = 1, \cdots, G$ represents the time elapsed until failure, and $C \sim Weibull(\alpha_g, \beta_g^*)$ represents the censoring time. Therefore, $W = min(F, C)$ and δ is the censorship indicator, where $\delta = 1$ if $F < C$ and $\delta = 0$ otherwise.

2.3 Definition of simulation scenarios and parametric values

With the variables simulated in both processes as described in the previous sections, scenarios were used in the Monte Carlo simulation under the factor combinations described in Table 2.

The Monte Carlo simulation procedure is justified for simulating samples, which computationally control different scenarios (Table 2). These scenarios serve as instruments for investigating the performance of the joint model, generating empirical distributions of the parameters. From these distributions, it becomes possible to estimate coverage probabilities, providing a more robust and detailed view of the model's behavior under various scenarios.

The values of the parameters of the Weibull model were defined arbitrarily, $\alpha = 12$ and $\beta = 4$. The correlation between repeated measures in the longitudinal process was determined to be $\rho = 0.5$.

2.4 Fit of the models for the survival and longitudinal processes with inclusion of categorical covariates and estimation of the probabilities of cross-coverage

Given the longitudinal process, the multilevel model was fitted with a random intercept (θ_0) and four categorical covariates ($\theta_1, \cdots, \theta_4$) whose systematic components were defined by the linear predictor Eq. (10).

$$\eta_k = \boldsymbol{\theta} \cdot X_k + \varepsilon, \quad (10)$$

Table 2: Scenarios considered for the simulation of data with different percentages of censorship (P), number of groups (Ng) and number of measurements (Nmed).

Scenario	Q	Structure	Ng	Nmed
1	15	AR(1)	20	30
2	15	AR(1)	20	60
3	15	AR(1)	20	100
4	15	Uniform	20	30
5	15	Uniform	20	60
6	15	Uniform	20	100
7	15	AR(1)	50	30
8	15	AR(1)	50	60
9	15	AR(1)	50	100
10	15	Uniform	50	30
11	15	Uniform	50	60
12	15	Uniform	50	100
13	50	AR(1)	20	30
14	50	AR(1)	20	60
15	50	AR(1)	20	100
16	50	Uniform	20	30
17	50	Uniform	20	60
18	50	Uniform	20	100
19	50	AR(1)	50	30
20	50	AR(1)	50	60
21	50	AR(1)	50	100
22	50	Uniform	50	30
23	50	Uniform	50	60
24	50	Uniform	50	100

in which $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$, $k = 1, \cdots, M \cdot G$ and $\varepsilon \sim N(0, 1)$.

For the survival process, the Weibull model was considered.

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}, \quad (11)$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$.

Once the estimates $\hat{\theta}$ and $\hat{\alpha}$ were obtained, the asymptotic confidence intervals were computed with the nominal level $\gamma = 0.95$ considering the averages of the estimates of the parameters of the longitudinal and survival models, adjusted in 1000 Monte Carlo realizations Eq. (12).

$$IC(\theta_i, \gamma) = \hat{\theta}_i \pm 1.96\sqrt{\hat{Var}(\hat{\theta}_i)}, \quad (12)$$

$$IC(\alpha_i, \gamma) = \hat{\alpha}_i \pm 1.96\sqrt{\hat{Var}(\hat{\alpha}_i)}, \quad (13)$$

in which $i = 1, \cdots, T_R = 4$.

As a function of the intervals, the estimate of the cross-coverage probability was denoted by $\widehat{CCP}_{\hat{\alpha}_i}\left(\hat{\theta}_i\right)$ as the frequency of the number of estimates $\hat{\alpha}_i$ obtained with the fit of the survival model, which is contained in the interval $IC(\theta_i, \gamma)$. Similarly, $\widehat{CCP}_{\hat{\theta}_i}(\hat{\alpha}_i)$ was estimated through the frequency of estimates $\hat{\theta}_i$ obtained with the fit of the multilevel model, which is contained in the interval $IC(\alpha_i, \gamma)$.

Following the recommendations of Bradley (1978) and Algina et al. (2005) and maintaining the nominal 95% confidence level, the confidence interval of this probability is [0.925; 0.975].

To obtain the results, a script was prepared in R software, version 4.0.4, for each scenario (Table 2) (R Core Team, 2022).

3 Results and discussion

Before discussing the simulated results, it should be noted that the first parameter of the survival model α_1 can be confused with the intercept (θ) of the longitudinal model. For didactic purposes and better clarification, let us assume the relative risk of the joint model, defined below:

$$h_i\left(t|\mathcal{M}_i(t), w_i\right) = h_0 \exp\left\{\gamma^\top w_i + a m_i(t)\right\}. \quad (14)$$

Specifying h_0 by the distribution Weibull (μ, σ) results in the following expression.

$$\begin{aligned} h_i\left(t|\mathcal{M}_i(t), w_i\right) &= \frac{\sigma}{\mu^\sigma} t^{\sigma-1} \exp\left\{\gamma^\top w_i + a m_i(t)\right\} \\ &= \frac{\sigma}{\exp\left\{\sigma \log(\mu)\right\}} t^{\sigma-1} \exp\left\{\gamma^\top w_i + a m_i(t)\right\} \\ &= \sigma t^{\sigma-1} \exp\left\{-\sigma \log(\mu) + \gamma^\top w_i + a m_i(t)\right\}. \end{aligned} \quad (15)$$

Therefore, the term $-\sigma \log(\mu)$ is conflated with the intercept effect of the linear predictor of the longitudinal model.

Given this understanding, although no inferences were made about relative risk, the results described in Table 3 correspond to the estimates of the cross-coverage probabilities. These are computed by the estimates of the first parameter of the survival model $\hat{\alpha}_i$ based on the confidence interval for the intercept of the longitudinal model.

Table 3: Probability of cross-coverage of the parameter α_1 of the survival model in relation to the intercept interval estimates θ_0 specified in the longitudinal model.

Scenario	P	Structure	$\widehat{CCP}_{\hat{\alpha}_1}\left(\hat{\theta}_0\right)$
1	15%	AR1	0.9586
4		Uniform	0.9513
9		AR1	0.9435
12		Uniform	0.9477
13	50%	AR1	0.9660
16		Uniform	0.9639
21		AR1	0.9568
24		Uniform	0.9439

The results described in Table 3 demonstrate that regardless of the correlation structure or whether the censorship percentage is specified as 15% or 50%, the estimates of the coverage probabilities approximately relate to the nominal confidence level defined in 95%.

The other scenarios involve the results of the other parameters of the Weibull and multilevel model in relation to the estimates of the cross-coverage probabilities. The extreme case, i.e., fewer groups (Ng=20) and fewer measurements (Nmed=30), follows the graphs illustrated in Fig. 1.

The results described in Fig. 1 demonstrate that, in general, the correlation structure to which the data are correlated has an impact. In this context, the estimates of the cross-coverage probabilities in both models showed discrepancies in at least one of the parameters in under the 95% nominal confidence level, with greater discrepancies when the percentage of censoring was high (50%).

This result did not occur under the uniform correlation structure, as shown in Figure 1(b)-1(d); that is, in both models, the longitudinal and survival processes were connected. In comparison with the results obtained by Villegas et al. (2013) and in relation to other survival models, the effect of correlation and percentages of censoring, different multiple failure approaches applying the Cox proportional hazards model are considered in different simulation scenarios.

Thus, assuming sample sizes fixed in n = (50,100,200,400); percentages of censorship p = 0%, 15%, 30% and 50%; number of recurring events $K = (3, 6, 9, 12)$; and the levels of correlation between the adjacent recurrence times fixed in $\rho = (0, 0.10, 0.45, 0.80)$, and without specifying the correlation structure, the authors concluded that the different approaches are stable against censorship and share a bias as the values increase For K recurrence levels, resulting in asymptotic confidence intervals that are imprecise relative to the specified nominal confidence level.

Fig. 2 illustrates the results of increasing the number of groups (Ng=50) and number of measurements (Nmeg=100). These results show that, given the correlation structure AR(1) (Fig. 2(a)-(c)) and for low censoring proportions, the estimates were reduced. This primarily caused an estimate in one of the parameters to be lower than the specification limit defined at 0.92 for the nominal confidence level. Thus, it has a limited ability to propose any recommendation regarding the existence of some connection between the longitudinal process and the survival process.

Under the uniform correlation, increasing the number of measurements resulted in reduced coverage probabilities in the presence of a high percentage of censorship (50), creating an estimate that is incoherent at the nominal confidence level in at least one of the model parameters of survival.

Štajduhar and Dalbelo-Bašić (2010) adapted the learning algorithms of Bayesian networks using a censorship weighting procedure proposed by Zupan et al. (2000), assuming nine different percentages of censoring p = 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80% and comparing the estimates of the Cox regression model. In this context, they concluded that the weighting

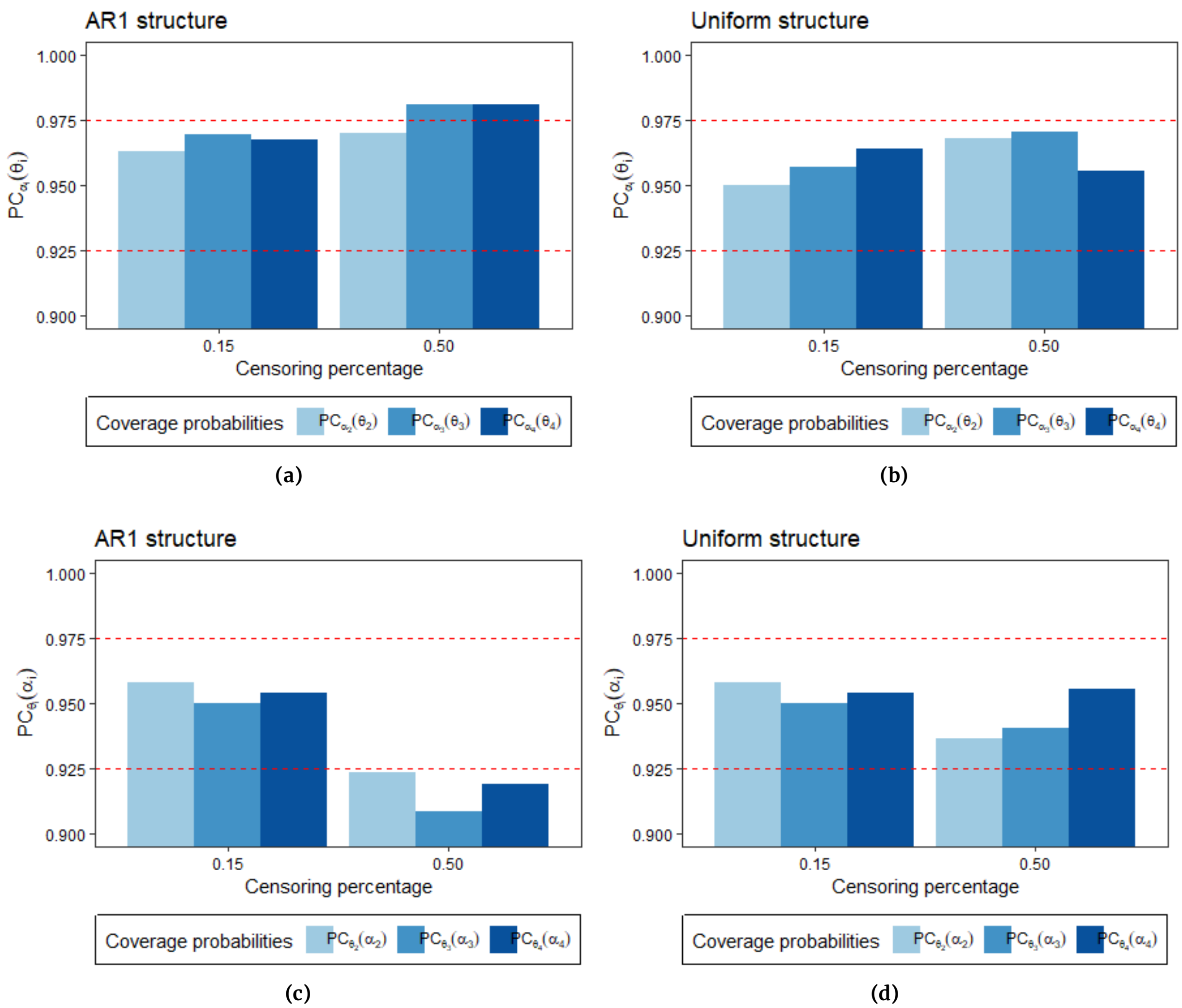

Figure 1: Estimates of the probabilities of cross-coverage, fixing the AR(1) and uniform correlation structures and censoring percentages of 0.15 and 0.50: (a) and (b), parameters of the longitudinal model in relation to the confidence interval of the parameters of the (c) and (d) parameters of the survival model in relation to the confidence interval of the parameters of the longitudinal model.

procedure should be used with Bayesian networks only with intermediate data censorship (from 40% to 60%). If data censorship is light (up to 30%), the original algorithms should be used.

Lin et al. (2013) performed a simulation study comparing the performance of several maximum likelihood estimation (ML) methods, the log-probit regression method and the nonparametric Kaplan–Meier method (KM). Thus, samples were generated from the following distributions: log-normal, gamma, a mixture of two log-normals and log-normal with 30% of observations at zero for different sample sizes. For each distribution evaluated, the percentage of censored observations was randomly generated from a uniform distribution ranging from 20% to 80%.

With these specifications, the results showed that the sample size had little impact on the accuracy of the estimates; however, the percentage of censored samples had the greatest impact, which is comparable to the results obtained by Antweiler and Taylor (2008) in the comparison of the maximum likelihood estimation methods, regression statistics by order and nonparameters for the analysis of left censored data; they concluded that with high percentages of censored data, the interval estimates were imprecise in relation to the nominal confidence level.

4 Conclusions

The proposed procedure used to estimate cross-coverage probabilities as a diagnostic tool for the connection of the longitudinal and survival models, was adequate when considering the Weibull model. Therefore, it can help researchers estimate a joint model that involves both

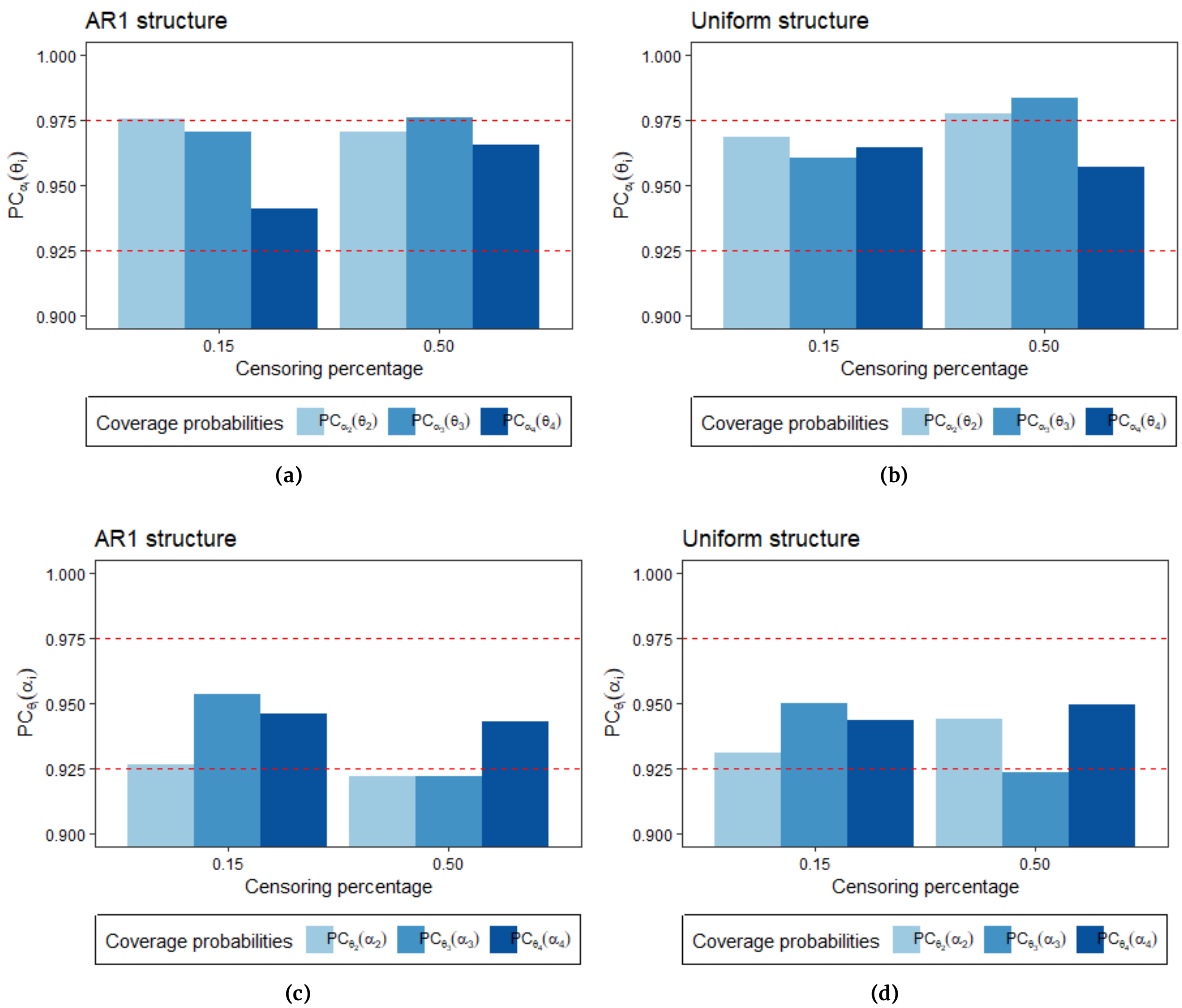

Figure 2: Estimates of the probabilities of cross-coverage, fixing the AR(1) and uniform correlation structures and censoring percentages of 0.15 and 0.50: (a) and (b), parameters of the longitudinal model in relation to the confidence interval of the parameters of the (c) and (d) parameters of the survival model in relation to the confidence interval of the parameters of the longitudinal model.

processes and minimize possible numerical convergence problems.

In the context of scenario 13, which includes smaller numbers of groups and measurements involving the AR1 correlation structure, the estimates of the cross-coverage probabilities in both models showed discrepancies in at least one of the parameters when the percentage rate of censorship was 50%, given a specified 95% nominal confidence level. Thus, this rate of censorship presents more harmful results than other rates. Under the same context but with a uniform correlation structure, as in scenarios 4 and 16, it is noted that a favorable condition exists for numerical convergence in obtaining maximum likelihood estimates for joint models of longitudinal and survival data.

Given the AR(1) correlation structure, increasing the number of groups and measurements and considering low censoring proportions leads to an estimate below the specification limit, which is defined as 0.92 for the nominal confidence level. This finding limits the ability to recommend the utilization of joint models for longitudinal and survival data.

For both correlation structures, increasing the number of measurements resulted in a reduction in the coverage probabilities in the presence of a high percentage of censorship, causing an estimate that was incoherent at the nominal confidence level in at least one of the parameters of the survival model. Thus, increasing the percentage of censorship negatively impacted the numerical convergence for obtaining maximum likelihood estimates of joint models for longitudinal and survival data.

The proposed methodology represents a significant advancement by offering a way to identify circumstances conducive to numerical convergence, which is essential for achieving results in joint modeling. The use of Monte Carlo simulation involving various levels of censoring and correlation structures provides a robust and comprehensive analysis, allowing for the evaluation of different scenarios and validation of the proposed methodology. Ultimately, this approach helps minimize computational challenges and convergence issues associated with joint models, thereby expanding their applicability across various fields.

In future studies, we aim to expand the methodology by considering additional parametric models. This could provide a broader view of scenarios where convergence issues may arise and evaluate the effectiveness of cross-coverage probability in these contexts. The findings of this study may contribute to the future development of computational tools incorporating the proposed methodology, assisting researchers in facilitating the estimation of joint models.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Algina, J., Keselman, H. and Penfield, R. D. (2005). An alternative to cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case., *Psychological methods* **10**(3): 317. https://psycnet.apa.org/doi/10.1037/1082-989X.10.3.317.

Antweiler, R. C. and Taylor, H. E. (2008). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. summary statistics, *Environmental science & technology* **42**(10): 3732–3738. https://doi.org/10.1021/es071301c.

Bradley, J. V. (1978). Robustness?, *British Journal of Mathematical & Statistical Psychology* **31**(2): 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x.

Colosimo, E. A. and Giolo, S. R. (2006). *Análise de sobrevivência aplicada*, Editora Blucher.

Do Ha, I. and Lee, Y. (2005). Multilevel mixed linear models for survival data, *Lifetime Data Analysis* **11**: 131–142. https://doi.org/10.1007/s10985-004-5644-2.

Giarola, L. T. P., Vivanco, M. J. F., Cirillo, M. A. and Menezes, F. S. (2018). Extended method for several dichotomous covariates to estimate the instantaneous risk function of the aalen additive model, *Journal of Modern Applied Statistical Methods* **17**(1): 27. https://doi.org/10.22237/jmasm/1543852660.

Hu, J. and Szymczak, S. (2023). A review on longitudinal data analysis with random forest, *Briefings in Bioinformatics* **24**(2): bbad002. https://doi.org/10.1093/bib/bbad002.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22. https://doi.org/10.1093/biomet/73.1.13.

Lin, N. X., Logan, S. and Henley, W. E. (2013). Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates, *Biometrics* **69**(4): 850–860. https://doi.org/10.1111/biom.12096.

Murray, J. and Philipson, P. (2022). A fast approximate em algorithm for joint models of survival and multivariate longitudinal data, *Computational Statistics & Data Analysis* **170**: 107438. https://doi.org/10.1016/j.csda.2022.107438.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*, Springer science & business media.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

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. https://doi.org/10.18637/jss.v035.i09.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*, CRC press.

Rizopoulos, D. and Lesaffre, E. (2014). Introduction to the special issue on joint modelling techniques, *Statistical methods in medical research* **23**(1): 3–10. https://doi.org/10.1177/0962280212445800.

Silva, J. A. D. and Cirillo, M. A. (2018). Selection criterion of work matrix as a function of limiting estimates of the covariance matrix of correlated data in gee, *Biometrical Journal* **60**(5): 979–990. https://doi.org/10.1002/bimj.201800035.

Štajduhar, I. and Dalbelo-Bašić, B. (2010). Learning bayesian networks from survival data using weighting censored instances, *Journal of biomedical informatics* **43**(4): 613–622. https://doi.org/10.1016/j.jbi.2010.03.005.

Sutradhar, B. C. and Das, K. (2000). On the accuracy of efficiency of estimating equation approach, *Biometrics* **56**(2): 622–625. https://doi.org/10.1111/j.0006-341X.2000.00622.x.

Verbeke, G., Molenberghs, G. and Verbeke, G. (1997). *Linear mixed models for longitudinal data*, Springer.

Villegas, R., Julià, O. and Ocaña, J. (2013). Empirical study of correlated survival times for recurrent events with proportional hazards margins and the effect of correlation and censoring, *BMC medical research methodology* **13**: 1–10. https://doi.org/10.1186/1471-2288-13-95.

Viviani, S., Alfó, M. and Rizopoulos, D. (2014). Generalized linear mixed joint model for longitudinal and survival outcomes, *Statistics and Computing* **24**: 417–427. https://doi.org/10.1007/s11222-013-9378-4.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics* pp. 175–188. https://doi.org/10.2307/2531905.

Zhang, D., Chen, M.-H., Ibrahim, J. G., Boye, M. E., Wang, P. and Shen, W. (2014). Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials, *Statistics in Medicine* **33**(27): 4715–4733. https://doi.org/10.1002/sim.6269.

Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R. and Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer, *Artificial intelligence in medicine* **20**(1): 59–75. https://doi.org/10.1016/S0933-3657(00)00053-1.