

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Cássius Henrique Xavier Oliveira

**A Class of Semiparametric Joint Frailty-Copula Models for Recurrent
Events Subjected to a Terminal Event.**

Belo Horizonte
2024

Cássius Henrique Xavier Oliveira

**A Class of Semiparametric Joint Frailty-Copula Models for Recurrent
Events Subjected to a Terminal Event.**

Final Version

Thesis presented to the Graduate Program in Statistics of the
Federal University of Minas Gerais in partial fulfillment of the
requirements for the degree of Doctor in Statistics.

Advisor: Fábio Nogueira Demarqui
Co-Advisor: Vinícius Diniz Mayrink

Belo Horizonte
2024

Oliveira, Cássius Henrique Xavier.

O48c A Class of semiparametric joint frailty-copula models for recurrent events subjected to a terminal event. [recurso eletrônico] / Cássius Henrique Xavier Oliveira. – 2024.
1 recurso online (98 f. il, color.) : pdf.

Orientador: Fábio Nogueira Demarqui.

Coorientador: Vinícius Diniz Mayrink.

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.

Referências: f.90-96

1. Estatística – Teses. 2. Análise de sobrevivência (Biometria) – Teses. 3. Correlação (Estatística) – Teses. 4. Modelo Yang and Prentice – Teses. 5. Bernstein, Polinômios de - Teses I. Demarqui, Fábio Nogueira. II. Mayrink, Vinícius Diniz. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV. Título.

CDU 519.2(043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA




FOLHA DE APROVAÇÃO

"A Class of Semiparametric Joint Frailty-Copula Models for Recurrent Events Subjected to a Terminal Event"


CÁSSIUS HENRIQUE XAVIER OLIVEIRA

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Doutor em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.


Aprovada em 12 de março de 2024, pela banca constituída pelos membros:

Documento assinado digitalmente
 **FABIO NOGUEIRA DEMARQUI**
Data: 14/03/2024 09:33:03-0300
Verifique em <https://validar.iti.gov.br>


Prof. Fábio Nogueira Demarqui – Orientador (DEST/UFMG)

Documento assinado digitalmente
 **VINICIUS DINIZ MAYRINK**
Data: 14/03/2024 09:33:31-0300
Verifique em <https://validar.iti.gov.br>


Prof. Vinícius Diniz Mayrink - Coorientador (DEST/UFMG)

Documento assinado digitalmente
 **ENRICO ANTONIO COLOSIMO**
Data: 13/03/2024 10:15:36-0300
Verifique em <https://validar.iti.gov.br>


Prof. Enrico Antonio Colosimo (DEST/UFMG)

Documento assinado digitalmente
 **MARIO DE CASTRO ANDRADE FILHO**
Data: 13/03/2024 09:40:41-0300
Verifique em <https://validar.iti.gov.br>

Prof. Mário de Castro Andrade Filho (ICMC/USP)

Documento assinado digitalmente
 **MARCOS OLIVEIRA PRATES**
Data: 13/03/2024 10:25:24-0300
Verifique em <https://validar.iti.gov.br>

Prof. Marcos Oliveira Prates (DEST/UFMG)

Documento assinado digitalmente
 **SILVANA SCHNEIDER**
Data: 12/03/2024 23:16:28-0300
Verifique em <https://validar.iti.gov.br>

Profª. Silvana Schneider (IME/UFRGS)

Belo Horizonte, 12 de março de 2024.

A minha família.

Agradecimentos

A Deus.

À Universidade Federal de Minas Gerais (UFMG), instituição de excelência e inovação, e aos estimados professores do Departamento de Estatística (DEST) que, com sua dedicação e paixão pelo ensino, me proporcionaram um conhecimento valioso e fonte de constante inspiração ao longo do doutorado.

Um agradecimento especial ao meu orientador, Professor Fábio Nogueira Demarqui, pelos ensinamentos decisivos, orientação consistente e exemplar profissionalismo que marcaram de forma indelével minha trajetória.

Ao coorientador, Professor Vinícius Diniz Mayrink, meu sincero agradecimento pelo auxílio inestimável e pelas contribuições significativas que enriqueceram imensamente este trabalho.

À Professora Marta Afonso Freitas, minha sincera gratidão por apoiar minha entrada no programa de doutorado e por ter sido uma orientadora tão dedicada durante meu mestrado. Seu encorajamento e sabedoria foram fundamentais para o meu crescimento profissional e pessoal.

À Professora Luciana Paula Reis, minha orientadora de graduação, cujo incentivo e suporte foram decisivos para que eu abraçasse o desafio da pós-graduação. Sua orientação foi um marco importante no início da minha jornada acadêmica.

À professora Maria Luíza Guerra de Toledo, estendo meus sinceros agradecimentos por todo o incentivo e apoio. Seus ensinamentos estatísticos tiveram um papel decisivo na definição de minha trajetória acadêmica.

Ao professor Marcelo Azevedo Costa, agradeço pelo incentivo contínuo, pelos ensinamentos e pela permissão para utilizar o laboratório LADEC, o que foi fundamental para o desenvolvimento de diversas etapas deste trabalho.

À Professora Vera Lúcia Souza, cujo papel foi fundamental e marcante em minha trajetória acadêmica, expresse minha profunda gratidão. Foi através de sua paixão e dedicação às ciências exatas que encontrei minha própria paixão por este campo, inspirando-me a seguir por este caminho. Seu dom e conhecimento foram contagiantes e desempenharam um papel crucial em moldar o amor e o interesse que hoje tenho pela área.

A todos os professores que tive ao longo da vida e que cumpriram com dedicação e excelência sua missão de ensinar, inspirando e me fazendo acreditar no poder transformador da educação.

Aos amigos, agradeço pelo suporte incondicional, pelas conversas enriquecedoras e

pelos momentos compartilhados que fortaleceram nossa amizade e suavizaram os desafios da vida acadêmica.

À minha família, especialmente à minha mãe Maria de Lourdes Xavier Oliveira e ao meu pai Gilmar Martins de Oliveira, minha tia Beth Xavier, meu primo Cauã Lucas, minha tia-avó Margarida Caio, minha avó Efigênia Caio (*in memoriam*), meu avô Jeremias Xavier (*in memoriam*) e minha tia-avó Raimunda Cosme (*in memoriam*), agradeço pelas orações, pelo incentivo constante e pelo amor incondicional que me sustentaram durante esta jornada.

À CAPES, FAPEMIG e CNPq expresso minha gratidão pelo suporte financeiro concedido durante o período do meu doutorado, fundamental para a realização deste estudo.

Por fim, a todos que, direta ou indiretamente, contribuíram para minha jornada acadêmica, meu mais sincero e profundo agradecimento! Vocês todos foram peças-chave nesta caminhada, e compartilho com cada um de vocês as conquistas alcançadas.

“Statistics is the grammar of science.”
(Karl Pearson)

Resumo

A análise de sobrevivência é uma das áreas mais importantes da estatística. Um de seus objetivos é avaliar potenciais fatores de risco na ocorrência de eventos. Os modelos de regressão de taxas de falha proporcionais (PH) são os recursos mais usados para esse fim, mas apresentam algumas limitações. Sua forte suposição de que a razão de taxas de falha é constante pode impedir o uso de modelos PH em algumas casuísticas. Alternativas ao modelo PH são discutidas na literatura, como os modelos de chances proporcionais (PO) e Yang e Prentice (YP). Entretanto, esses modelos não são capazes de acomodar a correlação entre eventos. Alguns trabalhos discutem a introdução de um efeito aleatório (ou fragilidade) na estrutura de regressão dos modelos PH e PO ou o uso de cópulas para acomodar dependências. Os dados de sobrevivência podem manifestar dependência de várias maneiras. O presente trabalho aborda casos em que um indivíduo pode vivenciar eventos sucessivos, chamados eventos recorrentes. Além disso, esses indivíduos estão sujeitos a experimentar um evento terminal, isto é, um evento que impede a continuidade do acompanhamento do indivíduo, não podendo, este, experimentar novos eventos recorrentes. Dessa forma, os processos de eventos recorrentes e terminal apresentam alguma dependência. Nosso objetivo é desenvolver, sob a abordagem Bayesiana, uma classe de modelos conjuntos de fragilidade-cópula para ajustar eventos recorrentes sujeitos a um evento terminal. Devido à forma matemática atrativa, usamos a cópula arquimediana de Clayton. Acoplamos polinômios de Bernstein (BP) e o modelo exponencial por partes (PEM) como funções de risco basais. Além disso, apresentamos uma classe de modelos de regressão Yang and Prentice para ajustar apenas os eventos terminais ou recorrentes usando as mesmas funções de linha de base. Apresentamos um estudo de simulação e exemplificamos nossos modelos através de uma aplicação.

Palavras-chave: sobrevivência; modelo Yang and Prentice; polinômios de Bernstein; fragilidade; cópulas.

Abstract

Survival analysis is one of the most important areas of statistics. One of its objectives is to assess potential risk factors in the occurrence of events. Proportional hazard (PH) regression models are the most commonly used tools for this purpose, but they have some limitations. Their strong assumption that the hazard rate ratio is constant can prevent the use of PH models in some cases. Alternatives to the PH model, such as proportional odds (PO) and Yang and Prentice (YP) models, are discussed in the literature. However, these models are not capable of accommodating the correlation between events. Some studies discuss the introduction of a random effect (or frailty) into the regression structure of the PH and PO models or the use of copulas to accommodate dependencies. Survival data can exhibit dependence in various ways. This work addresses cases where an individual may experience successive events, called recurrent events. Furthermore, these individuals are subject to experiencing a terminal event, that is, an event that prevents the continuation of the individual's follow-up, thus preventing new recurrent events. Therefore, the processes of recurrent and terminal events show some dependence. Our goal is to develop, under the Bayesian approach, a class of joint frailty-copula models to fit recurrent events subject to a terminal event. Due to its attractive mathematical form, we use the Archimedean Clayton copula. We couple Bernstein polynomials (BP) and the piecewise exponential model (PEM) as baseline hazard functions. Additionally, we present a class of Yang and Prentice regression models to fit only terminal or recurrent events using the same baseline functions. We present a simulation study and exemplify our models using a real case.

Keywords: survival; Yang and Prentice model; Bernstein polynomials; frailty; copulas.

List of Figures

1.1	Follow-up of an individual who experiences recurrent events and a terminal event.	20
1.2	Schematic representation of the model class proposed in this work.	22
1.3	Elucidation of the application of frailty and copula approaches on the occurrence of recurrent and terminal events of an individual.	24
2.1	Examples of survival curves when (a) $\psi = 1$ and $\phi = -1$, (b) $\psi = -1$ and $\phi = 1$, (c) $\psi = \phi = 0.5$, and (d) $\psi = 0$ and $\phi = 0.5$, in YP model.	31
2.2	The Bernstein basis functions of degree 7 on t in $[0, \tau]$	35
2.3	Example of $h(t)$ defined by PE.	37
2.4	Clayton copula, when $\theta = -0.9$: (A) density function, (B) cumulative distribution function, and (C) scatter plot. Clayton copula, when $\theta = 1$: (D) density function, (E) cumulative distribution function, and (F) scatter plot. Clayton copula, when $\theta = 20$: (G) density function, (H) cumulative distribution function, and (I) scatter plot.	42
3.1	Schematic representation of clustered survival times.	46
3.2	Schematic representation of data with recurrent events.	47
3.3	Schematic representation of multivariate survival times.	49
4.1	Schematic representation of the generation of the times to event considering individual frailties (class 1)	56
4.2	Schematic representation of the generation of the gap times between recurrent events considering shared frailties (class 1)	57
4.3	Boxplot of RB(%) for the YP_{EX} , YP_{PE} , and YP_{BP} models with individual frailties, for $L = 300$ and $M_C = 250$	59
4.4	Boxplot of RB(%) for the YP_{EX} , YP_{PE} , and YP_{BP} models with shared frailties, for $L = 300$ and $M_C = 250$	61
4.5	Schematic representation of the generation of times until the terminal event and the gap times between recurrent events using the Clayton copula (class 2)	63
4.6	Boxplot of RB(%) for the joint frailty-copula models: PH_{EX} , PH_{PE} when the generator is equivalent to the PH_{EX} model ($L = 300$ and $M_C = 250$).	68
4.7	Boxplot of RB(%) for the joint frailty-copula models: PO_{EX} , PO_{PE} when the generator is equivalent to the PO_{EX} ($L = 300$ and $M_C = 250$).	69

4.8	Boxplot of RB(%) for the joint frailty-copula models: YP_{EX} , YP_{PE} when the generator is equivalent to the YP_{EX} ($L = 300$ and $M_C = 250$)	70
5.1	Proportion of the number of readmissions.	72
5.2	Proportion of the categories of the covariates (A) sex, (B) chemotherapy treatment, and (C) Dukes's stage.	73
5.3	MCMC applied to YP_{BP} model: (A) Trace plots for the posterior samples; (B) Posterior density plots.	76
5.4	Kaplan-Meier (step function) and survival curves estimated by YP_{BP} model (continuous function) about the terminal event for the levels of variables (A) sex, (B) chemo, and (C) dukes. Time is measured in days.	77
5.5	MCMC applied to YP_{BP} model: (A) posterior trace plots for the posterior samples; (B) posterior density plots.	81
A.1	Numerical and graphical results of the Monte Carlo simulation study of the models of the first class of models.	97
A.2	Numerical and graphical results of the Monte Carlo simulation study of the models of the second class of models.	97
A.3	Numerical and graphical results of the real application of the models of the first class of models.	98
A.4	Numerical and graphical results of the real application of the models of the second class of models.	98

List of Tables

4.1	True values	57
4.2	Monte Carlo summary statistics of the YP_{EX} , YP_{PE} , and YP_{BP} models with individual frailties, for $L = 300$ and $M_C = 250$	58
4.3	Monte Carlo summary statistics of the YP_{EX} , YP_{PE} , and YP_{BP} models with shared frailties, for $L = 300$ and $M_C = 250$	60
4.4	True values	63
4.5	Monte Carlo summary statistics of the joint frailty-copula models: PH_{EX} , PH_{PE} , and PH_{BP} when the generator is equivalent to the PH_{EX} ($L = 300$ and $M_C = 250$).	65
4.6	Monte Carlo summary statistics of the joint frailty-copula models: PO_{EX} , PO_{PE} , and PO_{BP} when the generator is equivalent to the PO_{EX} ($L = 300$ and $M_C = 250$).	66
4.7	Monte Carlo summary statistics of the joint frailty-copula models: YP_{EX} , YP_{PE} , and YP_{BP} when the generator is equivalent to the YP_{EX} ($L = 300$ and $M_C = 250$).	67
5.1	Dummy variable for variable dukes	72
5.2	Summary of the PH and PO models fitted to the readmission data considering the terminal events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.	78
5.3	Summary of the YP models fitted to the readmission data considering the terminal events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.	79
5.4	Summary of the PH models fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.	82
5.5	Summary of the PO models fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.	83
5.6	Summary of the YP_{EX} model fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.	84

5.7	Summary of the YP_{PE} model fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.	85
5.8	Summary of the YP_{BP} model fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.	86

List of Symbols

T	Follow-up time to event
D	Follow-up time to death (terminal event)
R	Gap-times between the recurrent events or between the last recurrent event and the terminal event
L	Number of clusters
C	Follow-up time to right-censoring
Y	Observed follow-up time
$S(t)$	Survival function
$h(t)$	Hazard function
$h_0(t)$	Baseline hazard function
$H(t)$	Cumulative hazard function
$H_0(t)$	Baseline cumulative hazard function
$\mathcal{R}(t)$	Odds function
$\mathcal{R}_0(t)$	Baseline odds function
OR	Odds ratio
z	Frailty
w	Natural logarithm of frailty
$B_m^{C^*}(t)$	Bernstein polynomial of degree m for the continuous function $C^*(t)$
$b_{k,m}(t)$	Base of the Bernstein polynomial
$f_\beta(t, a, b)$	Probability density function of a Beta distribution with parameters a and b evaluated at t
\mathbf{x}	Vector of covariates of the regression model
$\mathbf{g}_m(t)$	Vector of base functions in the Bernstein polynomial
$\mathbf{G}_m(t)$	Vector of cumulative base functions in Bernstein polynomial
M_C	Number of Monte Carlo replicas
β	Vector of regression coefficients in PH and PO models
φ	Vector of base function parameters in the Bernstein polynomial
δ	Failure state indicator
ψ	Short-term regression coefficients vector in YP model
θ	Copula association parameter
Θ	Set of the parameters model
ϕ	Long-term regression coefficients vector in YP model
κ	Precision of the frailty

λ	Vector of hazard functions in Piecewise exponential model
ν	Short-term hazard ratios in YP model
ξ	Long-term hazard ratios in YP model
ρ	Time grid in Piecewise exponential model
σ_w	Standard deviation of the frailty
τ	Maximum of time-to-event or time to right censorship
τ_κ	Kendall's tau
Υ	Copula generating function

Abreviation

PH	Proportional hazards model
PO	Proportional odds model
YP	Yang and Prentice model
PH _{BP}	Proportional hazards model with Bernstein polynomials baseline and frailty
PH _{EX}	Proportional hazards model with exponential baseline and frailty
PH _{PE}	Proportional hazards model with piecewise exponential baseline and frailty
PO _{BP}	Proportional odds model with Bernstein polynomials baseline and frailty
PO _{EX}	Proportional odds model with exponential baseline and frailty
PO _{PE}	Proportional odds model with piecewise exponential baseline and frailty
YP _{BP}	Yang and Prentice model with Bernstein polynomials baseline and frailty
YP _{EX}	Yang and Prentice model with exponential baseline and frailty
YP _{PE}	Yang and Prentice model with piecewise exponential baseline and frailty
RB	Relative bias
ASE	Average standard error
SDE	Standard deviation estimate
LW	Lower bound of the credible interval
UP	Upper bound of the credible interval
CP	Coverage probability
par	Model parameters
est	Posterior average of model estimates
CI	Credible interval

Contents

1	Introduction	19
2	Survival analysis fundamentals	25
2.1	Introduction	25
2.2	Regression models	27
2.2.1	Proportional hazards model	27
2.2.2	Proportional odds model	28
2.2.3	Yang and Prentice model	29
2.3	Frailty model	30
2.4	Bernstein polynomials	33
2.5	Piecewise exponential model	36
2.6	Clayton copula	37
3	Proposed models	44
3.1	Class 1: Yang and Prentice frailty model	44
3.1.1	Notation and the likelihood function for clustered data	45
3.1.2	Notation and the likelihood function for data with recurrent events	47
3.2	Class 2: The joint frailty-copula models	48
3.2.1	Notation	48
3.2.2	The likelihood function	50
4	Monte Carlo simulation study	53
4.1	Analysis of the Yang and Prentice frailty models (Class 1)	56
4.1.1	Yang and Prentice model with individual frailty	58
4.1.2	Yang and Prentice model with shared frailty	59
4.2	Analysis of the joint frailty-copula models (Class 2)	61
5	Data analysis	71
5.1	Analysis of the Yang and Prentice frailty model (Class 1)	74
5.2	Analysis of the joint frailty-copula models (Class 2)	77
6	Final remarks and future research	87
	References	90

Appendix A Numerical and graphical results of all models	97
A.1 Monte Carlo simulation study	97
A.2 Real application	98

Chapter 1

Introduction

Survival analysis is a fundamental field in statistics that has witnessed significant advancements, particularly in the 1980s and 1990s (Colosimo and Giolo, 2006; Klein and Moeschberger, 2006). The primary focus of survival analysis is to study the time until the occurrence of an event of interest, such as the death of a patient or the failure of a mechanical equipment. A defining characteristic of survival data is the presence of censoring, which refers to incomplete information about the exact time when an event occurred.

In survival analysis, individuals are monitored over a specific period, and when the event of interest happens, its time is recorded. It is commonly assumed that the times until the event of interest occurrence are mutually independent. However, this assumption may not hold in certain scenarios. Consider situations where we want to evaluate the survival time of different litters of cats, the time until the onset of a disease in groups of twins, or the lifespan of patients treated in the same intensive care units. Assuming independence within these groups may be inappropriate, as individuals within the same group may exhibit similarities in the time until the event, which would not be observed in individuals outside these groups. It is important to note that these groupings can be either natural or artificial.

Individuals may also experience multiple occurrences of the same event, referred to as recurrent events. Examples include patients being infected by a virus multiple times or experiencing successive heart attacks, electrical systems encountering repeated failures in transmitting electricity, or the occurrence of repeated crimes in a specific area. In those cases, individuals can be interpreted as a group, where the “individual” represents the patient in the first and second examples, the electrical system in the third example, and the geographical areas in the last example.

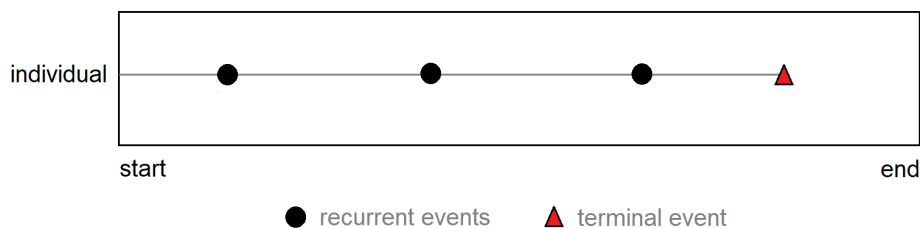
Often, these events are followed up until the individual becomes unavailable. There can be various reasons for unavailability, such as the individual moving to a different city or the patient’s decision not to proceed with an experimental treatment because of strong side effects. In these cases, the study is terminated without the individual experiencing the event of interest. However, there may be outcomes that lead to the discontinuation of follow-up, such as death. In such cases, we refer to the event that terminates the follow-up as a terminal event, as it occurs only once.

A very common goal in studies involving survival analysis is the evaluation of potential risk factors on the occurrence of events. Proportional hazards (PH) regression models, such as [Cox \(1972\)](#), are the most used approaches for this purpose. These models allow an intuitive interpretation of the regression parameters but have some limitations. Among them is the assumption of the hazard functions ratio of the observed individuals remaining constant over time. Another limitation is that PH models do not allow for accommodating potential correlation between recurrent events ([Amorim and Cai, 2015](#); [Li et al., 2019](#)).

It is a fact that recurrent events can have some kind of association. These events are very useful for assessing the deterioration of an individual's health status, as argued by [Huang and Wang \(2004\)](#). When survival times have an association induced by clusters or recurrences of events, we say that the data are multivariate. On the other hand, if the independence between times is not violated, we say that the survival data are univariate ([Colosimo and Giolo, 2006](#)).

[Colosimo and Giolo \(2006\)](#), [Hanagal \(2011\)](#) and others argue that a commonly used approach to deal with some dependence on survival data is to assume that these data have independence, conditioned on a set of unobserved variables, called frailty. The concept of frailty, introduced by [Vaupel et al. \(1979\)](#), defines it as a latent and multiplicative random variable. The authors used frailties, also called the random effect, to explain the effect of unobserved heterogeneity on the mortality of a population. [Clayton and Cuzick \(1985\)](#) used frailties to explain the heterogeneity about the hazard function in an extension of the PH model for multivariate survival data. [Huang and Wang \(2004\)](#) proposed a subject-level shared frailty model to accommodate the association between recurrent and terminal events. The name shared frailty is justified by the fact that each individual shares the same random effect on the hazard function of the terminal and recurrent events. [Figure 1.1](#) illustrates the follow-up of an individual who experiences some recurrent events (circle) and a terminal event (triangle).

Figure 1.1: Follow-up of an individual who experiences recurrent events and a terminal event.



Source: Prepared by the author.

In addition to frailty, there are studies that model the association between survival data through copulas. [Clayton \(1978\)](#) developed a survival model for bivariate time-to-

event data using copulas. Other copula applications can be seen in [Oakes \(1982\)](#), [Suzuki \(2012\)](#), [Biondo and Suzuki \(2016\)](#), [Prenen et al. \(2017\)](#) and [Patiño \(2018\)](#). [Emura et al. \(2017\)](#) developed a joint frailty-copula model as an extension to the shared frailty model for meta-analysis. [Li et al. \(2019\)](#) introduced a joint frailty-copula model, in which the random effect explains the correlation between the recurrent events of an individual and a copula is used to model the association between the recurrent and the terminal events of each individual.

As for the regression models, there are alternatives to the PH model. The proportional odds (PO) model is one of them and was introduced by [Bennett \(1983\)](#). The hypothesis of this model is that the survival curves approximate each other, but do not intersect ([Collett, 2015](#)). Although these models also cannot explain the correlation between the data, it is possible to incorporate on them a random effect for this purpose ([Economou and Caroni, 2007](#)). Another alternative is the Yang and Prentice (YP) regression model introduced by [Yang and Prentice \(2005\)](#), which includes the PH and PO models as particular cases. Here, the survival functions are allowed to intersect, and this provides an advantage over the PH and PO models ([Demarqui and Mayrink, 2021](#)).

The arguments presented evidence that researchers are seeking to develop even more realistic and, consequently, more complex models. It is known that computational advances allowed the development of classical and Bayesian methodologies in survival analysis ([Ibrahim et al., 2014](#)). Programming languages such as R ([R Core Team, 2024](#)), for example, allow for the implementation of inferential methods. Several researchers have been developing computational packages that facilitate the replication of published results and promote a greater flow of knowledge and usability of their methods. In this work, we will focus on using the R language.

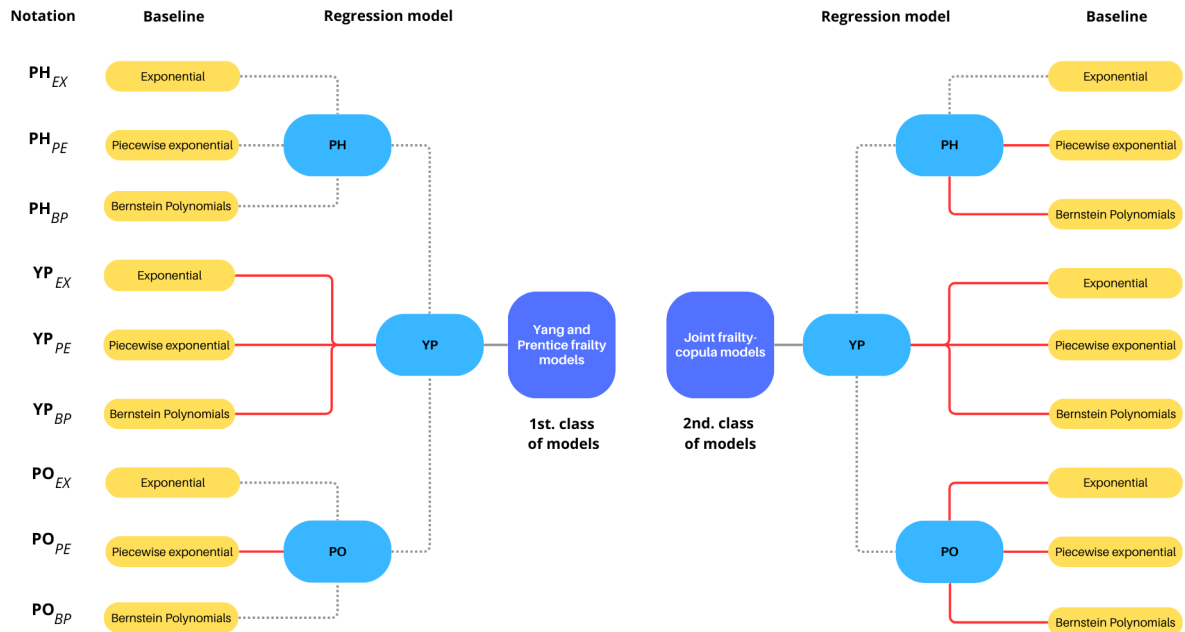
One of the aims of this work is to present two classes of models, under the Bayesian approach, to explain the effect of observed characteristics on the process of recurrent events that can be interrupted by a terminal event or a loss of follow-up due to external factors. We chose to use copula with the shared frailty model to have a clear and directly interpretable measurement of the association between recurrent and terminal events, as argued by [Li et al. \(2019\)](#). We further choose to use the YP regression structure since it generalizes the PH and PO models.

We will also be able to assess the potential risk characteristics of individuals into clusters using shared frailty models in a YP regression structure. Furthermore, we will be able to study univariate survival data also using YP regression models.

To clarify the contributions of this thesis, see [Figure 1.2](#). The notation designated for the models is presented within the same figure, specifically in the row associated with the corresponding models. In this figure, dotted lines denote models that are already established in the literature and can be considered as special cases of the models proposed in this work. The solid red lines represent the original models developed in this thesis.

Notably, we introduce novelty in the frailty models with a PO regression structure when we combine it with a piecewise exponential baseline, a concept not previously found in the literature.

Figure 1.2: Schematic representation of the model class proposed in this work.



Source: Prepared by the author.

Our models for survival data encompassing both recurrent and terminal events also stand as significant contributions to the field. We have developed PH models that utilize piecewise exponential and Bernstein baselines at the same time that we use the Clayton copula for modeling the associations between recurrent and terminal events, as well as frailties to model the association among recurrent events within the same individual. These approaches are not found in the current literature.

Additionally, this thesis presents innovations in PO and YP families of regression models: the use of exponential, piecewise exponential, and Bernstein baselines, along with the incorporation of the Clayton copula and frailties. Since we are not aware of any existing studies that have documented these approaches, the models proposed here signify advancements in the methodologies of survival analysis.

Goal

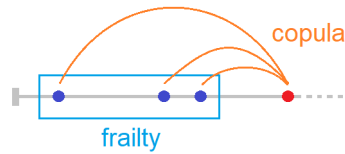
The primary goal of this study is to develop two classes of frailty models within the Bayesian framework. Figure 1.3 illustrates the structure of these classes.

- The first class of models will have a frailty YP with three baseline functions available: exponential, piecewise exponential, and Bernstein polynomial. It will allow to analyze survival data arranged in three configurations:
 - 1st.: The frailty term will serve to explain unobserved heterogeneities, that is, variations in survival time that are not explained by the fixed effects of the models. In this case, we will refer to the element of the frailty as individual frailty.
 - 2nd.: Individuals have only one time until the event but are arranged in artificial or natural clusters. Individuals in the same cluster will share the same frailty due to the similarities resulting from the grouping. Frailty (here referred to as shared frailty) will therefore serve to accommodate a possible association between the survival times of individuals belonging to the same cluster.
 - 3rd.: Individuals present recurrent events. Thus, frailty will be used to accommodate the association between the survival times of the same individual. In this context, we can understand the individual as a cluster and frailty here will also be referred to as shared frailty.
- The second class of models consist of three regression families: PH (proportional hazard), PO (proportional odds), and YP (Yang and Prentice) allowing great flexibility. Regression structures will have three possible baseline functions: exponential, piecewise exponential, and Bernstein polynomials. This class of models will allow us to analyze survival data in which individuals may experience recurrent events and a terminal event or administrative censoring. Loss of follow-up induces a dependent censoring of the individual's process of recurrent events. In our models, individual's recurrent events will share a term of frailty that will serve to accommodate a possible association between the gap times of these recurrences. Terminal events that may also have a dependence on recurrent events will be incorporated into the likelihood function using the Clayton copula. Figure 1.3 elucidates how interactions between the survival times of recurrent and terminal events will be handled within this model class.

The selection of the Clayton copula for this study is primarily due to its simplicity and widespread usage among Archimedean copulas. Additionally, our models draw inspi-

ration from the work of [Li et al. \(2019\)](#), who also employed the Clayton copula in their model.

Figure 1.3: Elucidation of the application of frailty and copula approaches on the occurrence of recurrent and terminal events of an individual.



Source: Prepared by the author.

Text structure

This thesis is organized as follows. Chapter 2 presents some fundamental concepts of survival analysis, the PH, PO, and YP models, as well as a description of the frailty model. The Bernstein polynomials and the piecewise exponential model, which will be used to handle the baseline hazard functions, are also presented in this chapter. Besides, it discusses concepts and properties of copulas and introduces the Clayton copula. Chapter 3 presents our proposed models. It starts by setting the notation and then explains the construction of the likelihood function. Chapter 4 discusses the data generation steps and the results of the Monte Carlo study. Chapter 5 presents a real application. We close this text with discussions of some results and perspectives for future research in Chapter 6.

Chapter 2

Survival analysis fundamentals

2.1 Introduction

In survival analysis, the response variable is the time until the occurrence of an event of interest, called the failure time. The main characteristic of survival data is the presence of censoring which is an incomplete observation. It can occur for several reasons that can be or not related to the study (Klein and Moeschberger, 2006; Schneider, 2017). If the study ends before the individual experiences the event, the censoring mechanism is administrative. Another type of censoring is called dropout, in which the individual leaves the study for external reasons. If, however, this loss of follow-up is associated with the study, we say that the censoring mechanism is informative. An example of this type of censoring is when an individual leaves the study due to side effects from the treatment. When it happens, it is expected that there is a correlation between time-to-event and time-to-censoring, and ignoring this possible correlation can cause biased estimates (Huang and Wolfe, 2002; Schneider et al., 2020).

Another case in which censoring is said informative occurs when the individual experiences recurrent events and a terminal event (Huang and Wang, 2004; Huang and Liu, 2007; Li et al., 2019). The terminal event might be related to the recurrent events experienced by that individual. This is because the terminal event prevents the continuity of the individual's follow-up regarding recurrent events.

As for the types of censoring, we can establish other classifications: right-censoring, left-censoring, and interval-censoring. Right-censoring occurs when the follow-up time is not enough for the individual to experience the event of interest. It can be a type I censoring - when the study duration is specified in advance and, therefore, the number of events is random. A type II censoring - when the number of failures is defined before the beginning of the follow-up. A random censoring is observed when an individual leaves the study for a reason not related to it. The left-censoring occurs when the individual begins to be accompanied, having already experienced the event of interest at some unknown moment in the past. Finally, interval-censoring is defined when an individual experiences

an event at an unknown moment between two observed times. In this work, we will focus on right-censoring. The data used are right-censored, either due to a loss of follow-up of the individual for reasons outside the study or due to the occurrence of a terminal event that prevents new recurrent events.

Let $T \geq 0$ be the random variable denoting the time-to-event of an individual and let $C \geq 0$ be the random variable representing the time-to-censoring for the same individual. An observation is right-censored when $T > C$. In this case, the time of observation of the individual is

$$Y = \begin{cases} T, & \text{if } T \leq C; \\ C, & \text{otherwise.} \end{cases}$$

Let δ be an indicator such that $\delta = 1$ indicates that the observed time of an individual is a time-to-event or, mathematically, $\delta = I(T \leq C)$.

Now, let's discuss some important functions in survival analysis. One of them is the survival function. It is defined as the probability that an individual does not experience an event until a certain time t , that is, the probability that he or she will survive until t (Colosimo and Giolo, 2006; Klein and Moeschberger, 2006). Denote by $F(t) = P(T \leq t)$ the cumulative distribution function (c.d.f) of the random variable T , that is, the probability of an individual experience an event up to time t . In this way, the survival function and the cumulative distribution function are complementary, and therefore, the survival function is given by

$$S(t) = 1 - F(t).$$

Another relevant function is called the hazard function. We denote it by $h(t)$. This function is more informative than the survival function since similar survival functions can have different hazard functions (Colosimo and Giolo, 2006). The hazard function is given by

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t},$$

Furthermore, the cumulative hazard function is given by

$$H(t) = \int_0^t h(u) du.$$

In this work, we will use the odds function $\mathcal{R}(t)$. It is defined as the ratio between the probability of an individual experiencing an event until a time t and the probability of surviving until that time (Bennett, 1983). Consider:

$$\mathcal{R}(t) = \frac{F(t)}{1 - F(t)} = \frac{F(t)}{S(t)}.$$

Some well-known relationships between the mentioned functions are as follows:

$$\begin{aligned} f(t) &= -\frac{d}{dt}S(t) = h(t) \exp[-H(t)], \\ H(t) &= -\log S(t) \iff S(t) = \exp[-H(t)], \\ h(t) &= \frac{d}{dt}H(t) = -\frac{S'(t)}{S(t)}, \text{ with } S'(t) = \frac{d}{dt}S(t), \end{aligned}$$

and

$$\mathcal{R}(t) = \exp[H(t)] - 1.$$

See additional details in [Hosmer and Lemeshow \(1999\)](#), [Klein and Moeschberger \(2006\)](#), [Kleinbaum and Klein \(2010\)](#), [Lawless \(2011\)](#), and [Ibrahim et al. \(2014\)](#). The next sections will discuss some important regression models in survival analysis.

2.2 Regression models

In this section, three important regression models will be presented. They are the PH model, the PO model, and the YP model, of which the PH and PO models are particular cases. For these models, denote by $\mathbf{x} = (x_1, \dots, x_p)$ a row vector of explanatory variables.

2.2.1 Proportional hazards model

The proportional hazards model, introduced by [Cox \(1972\)](#), allows to assess the effect of some characteristics of an individual on the time-to-event. These characteristics are incorporated in the hazard function. The PH model is one of the most used in clinical studies due to its versatility. Its hazard function is commonly given by ([Colosimo and Giolo, 2006](#); [Kalbfleisch and Prentice, 2011](#))

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}\boldsymbol{\beta}), \tag{2.1}$$

where the parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ are the regression coefficients and $h_0(\cdot)$ is a non-negative function called baseline hazard, which does not depend on \mathbf{x} , and can be modeled either parametrically or non-parametrically. The intercept β_0 is not explicit because it is incorporated by the term $h_0(t)$ ([Colosimo and Giolo, 2006](#)).

The PH model has the property of a constant ratio in time between the hazard functions of two different individuals, as long as the effect of covariates is invariant over time. Consider two individuals, i and j , then the ratio between their hazard functions is given by

$$\frac{h_i(t|\mathbf{x}_i)}{h_j(t|\mathbf{x}_j)} = \frac{h_0(t) \exp(\mathbf{x}_i\boldsymbol{\beta})}{h_0(t) \exp(\mathbf{x}_j\boldsymbol{\beta})} = \exp\{(\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\beta}\}.$$

In this sense, the parameters $\boldsymbol{\beta}$ will allow the identification of which covariates increase or decrease the hazard of an individual experiencing a failure.

The survival function for the PH model can be rewritten as

$$S(t|\mathbf{x}) = \exp\{-H_0(t) \exp(\mathbf{x}\boldsymbol{\beta})\},$$

where $\exp\{-H_0(t)\} = S_0(t)$ is the baseline survival function.

When the assumption of proportional hazards is violated, the model described in this section becomes inappropriate. Furthermore, when there is some dependence between the survival data, the PH model is also inadequate as it does not explain the correlation between events (Colosimo and Giolo, 2006; Li et al., 2019). This problem is easily solved by introducing frailty. In addition, when the censoring mechanism is informative, the PH model does not offer valid estimates (Schneider et al., 2020). There are several approaches in the literature that provide alternatives to the PH model. Two of those alternatives are described in the subsequent sections: PO and YP models. These models also depend on the assumption of independence of observations.

2.2.2 Proportional odds model

The proportional odds model was introduced by Bennett (1983) to deal with situations in which the survival curves become closer as $t \rightarrow \infty$, but do not intersect. We can cite some works that bring applications of this model as Royston and Parmar (2002), Hanson and Yang (2007), Wang and Dunson (2011), and Panaro (2020).

Let $\mathcal{R}(t)$ be the odds function and let $\mathcal{R}_0(t)$ be the baseline odds function, that is,

$$\mathcal{R}_0(t) = \frac{1 - S_0(t)}{S_0(t)}.$$

This model can be characterized as follows:

$$\mathcal{R}(t|\mathbf{x}) = \frac{F(t|\mathbf{x})}{S(t|\mathbf{x})} = \mathcal{R}_0(t) \exp(\mathbf{x}\boldsymbol{\beta}), \quad (2.2)$$

The survival function is given by

$$S(t|\mathbf{x}) = \frac{1}{1 + \mathcal{R}(t|\mathbf{x})},$$

and the cumulative hazard function is defined as

$$H(t|\mathbf{x}) = -\log[\mathcal{R}(t|\mathbf{x}) + 1].$$

It is possible to verify that the ratio between the odds functions of two individuals, i and j , is constant over time. This is an assumption of the PO model. One can write

$$OR = \frac{\mathcal{R}_0(t) \exp(\mathbf{x}_i \boldsymbol{\beta})}{\mathcal{R}_0(t) \exp(\mathbf{x}_j \boldsymbol{\beta})} = \exp\{(\mathbf{x}_i - \mathbf{x}_j) \boldsymbol{\beta}\}.$$

Some considerations can be made regarding the value of OR : if $OR = 1$, the failure is equally likely to happen for both individuals. If $OR > 1$, the individual i is more likely to experience failure compared to j ; and the opposite happens when $OR < 1$.

The PO model, as well as the PH model, can be understood as a particular case of the YP regression model (Yang and Prentice, 2005). The next section provides some details about the YP approach.

2.2.3 Yang and Prentice model

As already discussed, both the PH and the PO models are very important in survival analysis, but they cannot be applied in situations where the hazard ratio and odds ratio, respectively, are not constant. To deal with this limitation, Yang and Prentice (2005) proposed a model in which survival curves can intersect. This model can be characterized in terms of the survival function

$$S(t|\mathbf{x}) = \left[1 + \frac{\nu}{\xi} \mathcal{R}_0(t) \right]^{-\xi}, \quad (2.3)$$

where $\nu = \exp(\mathbf{x}\boldsymbol{\psi})$ and $\xi = \exp(\mathbf{x}\boldsymbol{\phi})$, $\boldsymbol{\psi} = (\psi_1, \dots, \psi_p)'$ and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$ are vectors of regression parameters without intercepts. We are evaluating the impact of the same variables over both short and long terms, although this is not required. The function \mathcal{R}_0 is the baseline odds as defined in Section 2.2.2.

We can express the hazard function of this model as

$$h(t|\mathbf{x}) = \frac{\nu \xi}{\nu F_0(t) + \xi S_0(t)} h_0(t), \quad (2.4)$$

where $F_0(t)$ is the baseline cumulative distribution function, $S_0(t) = 1 - F_0(t)$ is the baseline survival function and $h_0(t)$ is the baseline hazard function.

The YP model can be reduced to the PH and PO models. Note that when $\boldsymbol{\psi} = \boldsymbol{\phi}$,

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}\boldsymbol{\psi}) = h_0(t) \exp(\mathbf{x}\boldsymbol{\phi}),$$

and this is the hazard function of the PH model, as shown in Expression (2.1). If $\boldsymbol{\phi} = \mathbf{0}$, we have

$$S(t|\mathbf{x}) = [1 + \mathcal{R}_0(t) \exp(\mathbf{x}\boldsymbol{\psi})]^{-1} \Rightarrow \mathcal{R}(t|\mathbf{x}) = \frac{F(t|\mathbf{x})}{S(t|\mathbf{x})},$$

and this is the expression of the odds function in the PO model, as presented in (2.2). When $\psi_j \phi_j < 0$, for any pair of coefficient (ψ_j, ϕ_j) , with $j \in \{1, \dots, p\}$, the survival curves intersect.

Another feature of the YP model is related to the hazard ratio limits, if $t \rightarrow 0$ or $t \rightarrow \infty$. When $t \rightarrow 0$, we have

$$\lim_{t \rightarrow 0} \frac{h(t|\mathbf{x})}{h(t|\mathbf{0})} = \nu,$$

where ν is interpreted as the short-term hazard ratios, and $\boldsymbol{\psi}$ as the short-term regression coefficients vector. Additionally, if $t \rightarrow \infty$,

$$\lim_{t \rightarrow \infty} \frac{h(t|\mathbf{x})}{h(t|\mathbf{0})} = \xi.$$

We can interpret ξ as the long-term hazard ratios, and $\boldsymbol{\phi}$ as the long-term regression coefficients vector.

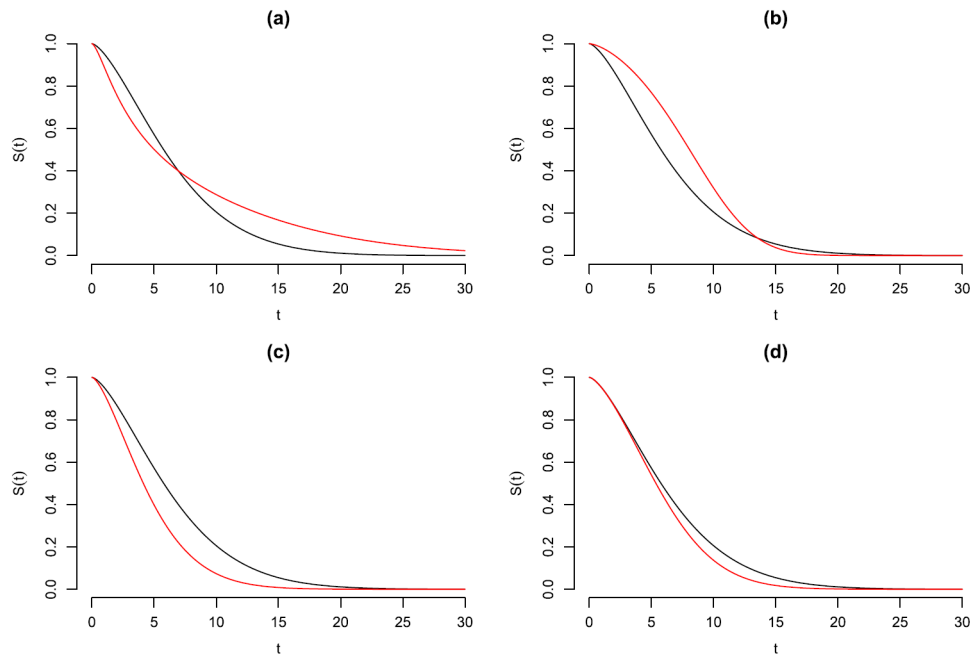
Demarqui and Mayrink (2021) illustrate four different scenarios generated by choosing some values of $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. It is possible to see the effect of these choices on the survival functions. They intersect when $\psi = 1$ and $\phi = -1$ and when $\psi = -1$ and $\phi = 1$. This is shown in Figures 2.1-(a) and 2.1-(b), respectively. When $\psi = \phi = 0.5$, the model reduces to the PH model, as can be seen in Figures 2.1-(c). Finally, choosing $\psi = 0$ and $\phi = 0.5$, one finds a structure of proportional odds shown in Figures 2.1-(d).

Several works address the YP model in the literature, as Yang and Zhao (2012), Diao et al. (2013), Wang (2013) and Demarqui et al. (2019). Demarqui and Mayrink (2021) proposed a semiparametric model for survival data using YP regression and the piecewise exponential as the baseline hazard function. The fit of this model can be done using by R package YPPE from the first author (Demarqui, 2020b). The inference is done under the frequentist and the Bayesian approaches.

2.3 Frailty model

In survival analysis, we are generally interested in identifying the factors that can increase or decrease an individual's hazard of experiencing the event of interest. However,

Figure 2.1: Examples of survival curves when (a) $\psi = 1$ and $\phi = -1$, (b) $\psi = -1$ and $\phi = 1$, (c) $\psi = \phi = 0.5$, and (d) $\psi = 0$ and $\phi = 0.5$, in YP model..



Source: [Demarqui and Mayrink \(2021\)](#).

not all characteristics of an individual are known or measurable. In the literature, it is common for such unknown factors to be defined as individual heterogeneity or frailty. [Clayton \(1978\)](#) introduced this concept to explain that different individuals may present different hazards even though their measurable attributes are similar. [Vaupel et al. \(1979\)](#) introduced the term “frailty” as a latent and multiplicative random variable on the mortality rate of individuals. This random variable, in that study, absorbs the unobserved heterogeneity.

Since the work of [Vaupel et al. \(1979\)](#), the frailty model has received a great deal of attention in the literature. [Clayton and Cuzick \(1985\)](#) proposed an extension of the proportional hazards model ([Cox, 1972](#)) to account for multivariate survival data by the addition of a random effect representing unobserved heterogeneity. Frailty models can also be used to accommodate the association between recurrent events, as in [Lawless \(1987\)](#). These models were further applied to handle recurrent events in the presence of a terminal event such as in [Huang and Wang \(2004\)](#), [Liu et al. \(2004\)](#) and [Mazroui et al. \(2012\)](#). [Schneider et al. \(2020\)](#) used the frailty to fit survival data subjected to dependent censoring.

Frailty models can also be used to accommodate the correlation between individuals belonging to the same group or cluster. In these cases, the model is called the shared frailty model, since individuals from the same group share the random effect. The shared frailty framework is understood as an extension of the PH model. Let $i = 1, \dots, n_k$ be the

individual i of the group k , with $k = 1, \dots, q$, and n_k be the number of individuals in the k -th group. The hazard function of this individual is given by:

$$\begin{aligned} h_{i,k}(t|x_{i,k}, z_k) &= h_0(t)z_k \exp(\mathbf{x}_{i,k}\boldsymbol{\beta}) \\ &= h_0(t) \exp(\mathbf{x}_{i,k}\boldsymbol{\beta} + w_k), \end{aligned}$$

where $z_k = \exp(w_k)$ is the frailty of the k -th group, $\mathbf{x}_{i,k} = (x_{i,k,1}, \dots, x_{i,k,p})$ is a row vector of covariates. Colosimo and Giolo (2006); Klein and Moeschberger (2006) highlight that w_k is usually assumed to have a distribution with zero mean and unknown variance. Other distributions can be chosen for z_k as the gamma, and positive stable distribution, for example; see Hougaard (2012) and Wienke (2020) for more details.

According to Colosimo and Giolo (2006), the presence of the random element in the PH model generates different interpretations for the hazard ratio.

1. When individuals i and j are from different groups $k \neq k'$, the hazard ratio is

$$\frac{h_{i,k}(t|x_{i,k}, z_k)}{h_{j,k'}(t|x_{j,k'}, z_{k'})} = \frac{h_0(t)z_k \exp(\mathbf{x}_{i,k}\boldsymbol{\beta})}{h_0(t)z_{k'} \exp(\mathbf{x}_{j,k'}\boldsymbol{\beta})} = \frac{z_k}{z_{k'}} \exp\{(\mathbf{x}_{i,k} - \mathbf{x}_{j,k'})\boldsymbol{\beta}\}.$$

Thus, the ratio between the hazard functions depends not only on the observed characteristics but also on the random effects of the two individuals.

2. Let i and j be individuals from the same group k . Both individuals have the same z_k element of frailty. In this case, the hazard ratio is given by

$$\frac{h_{i,k}(t|x_{i,k}, z_k)}{h_{j,k'}(t|x_{j,k'}, z_k)} = \exp\{(\mathbf{x}_{i,k} - \mathbf{x}_{j,k})\boldsymbol{\beta}\},$$

and the interpretation follows the PH model.

3. Now consider two individuals, i and j , who have equal values of covariates but are from different groups $k \neq k'$. The hazard ratio is

$$\frac{h_{i,k}(t|x_{i,k}, z_k)}{h_{j,k'}(t|x_{j,k'}, z_{k'})} = \frac{z_k}{z_{k'}}.$$

Here, the ratio between the hazard functions is the ratio of frailties.

We have discussed the incorporation of a random effect into the PH model. However, there are works that also use frailty in PO models as Economou and Caroni (2007), Lin and Wang (2011), and Gupta and Peng (2014). In this case, the frailty is inserted in the linear predictor shown in (2.2).

2.4 Bernstein polynomials

Polynomials are mathematical tools that have attractive features such as the fact that they can be easily derived and integrated. [Bernstein \(1912\)](#) introduced a polynomial that is a linear combination of basis and is known as the Bernstein polynomial. Let m be a positive integer, the Bernstein polynomials of degree m for the continuous function $C^*(t)$, defined in a range $[0, 1]$, can be written as

$$B_m^{C^*}(t) = \sum_{k=0}^m C^* \left(\frac{k}{m} \right) b_{k,m}(t); \quad t \in [0, 1], \quad (2.5)$$

where $b_{k,m}(t)$ is the basis of the polynomial such that

$$b_{k,m}(t) = \binom{m}{k} t^k (1-t)^{m-k}; \quad k \in 0, \dots, m.$$

The Bernstein polynomials became an important mathematical tool to prove the Weierstrass approximation theorem, which states that a continuous function in the closed interval $[a, b]$, with $a \in \mathbb{R}$ and $b \in \mathbb{R}$, $a < b$, can be approximated arbitrarily by a polynomial ([Lorentz, 1986](#)).

[Farouki and Rajan \(1987\)](#) present a formulation for the Bernstein polynomials to accommodate a continuous function $C^*(t)$ restricted to the closed interval $[a, b]$ as

$$B_m^{C^*}(t) = \sum_{k=0}^m C^* \left(a + \frac{k}{m}(b-a) \right) b_{k,m} \left(\frac{t-a}{b-a} \right); \quad t \in [a, b]. \quad (2.6)$$

[Feller \(1987\)](#) shows that $B_m^{C^*}(t)$ converges uniformly to $C^*(t)$, when $m \in \mathbb{N}$ is chosen arbitrarily greater than $M \in \mathbb{N}$. In other words,

$$\forall \epsilon > 0, \exists M \in \mathbb{N} \text{ such that } m > M \Rightarrow |B_m^{C^*}(t) - C^*(t)| < \epsilon, \forall t \in [0, 1].$$

The Bernstein polynomials in the form described in (2.5) can be rewritten as the expected value of the $C^* \left(\frac{K}{m} \right)$ with $K \sim \text{Binomial}(m, t)$. This is,

$$E \left[C^* \left(\frac{K}{m} \right) \right] = \sum_{k=0}^m C^* \left(\frac{k}{m} \right) \binom{m}{k} t^k (1-t)^{m-k}. \quad (2.7)$$

Note that (2.5) and (2.7) are equivalent ([Koralov and Sinai, 2007](#)). As we consider $K \sim \text{Binomial}(m, t)$, we have

$$\sum_{k=0}^m b_{k,m}(t) = 1.$$

[Farouki \(2012\)](#) presents some important properties of the Bernstein polynomial:

1. Non-negativity: the basis of the polynomial are non-negative for any $t \in [0, 1]$, that is,

$$\begin{cases} b_{k,m} \geq 0, & \text{if } 0 \leq k \leq m, \\ b_{k,m} \equiv 0, & \text{if } k < 0 \text{ or } k > m. \end{cases}$$

2. Symmetry: the basis of the polynomial $b_{k,m}$ and $b_{m-k,m}$ are symmetric in $t = \frac{1}{2}$, that is

$$b_{m-k,m}(1-t) = b_{k,m}(t).$$

Figure 2.2 shows the basis functions of a polynomial of degree 7 and illustrates this property. Each basis of the BP is represented by a curve.

3. Recursion: The basis of degree $m + 1$ can be generated from the basis of degree m using the relation

$$b_{k,m+1}(t) = tb_{k-1,m}(t) + (1-t)b_{k,m}(t),$$

for $k = 0, 1, \dots, m + 1$ and starting the recursion with $b_{0,0} \equiv 1$.

4. Derivatives: The basis of the polynomial satisfy the equation

$$\frac{d}{dt}b_{k,m}(t) = m [b_{k-1,m-1}(t) - b_{k,m-1}(t)].$$

Now, consider $t \in [0, \tau]$. The derivative of Bernstein polynomials with respect to t can be written as (Osman and Ghosh, 2012):

$$b_m^{C^*}(t) = \sum_{k=1}^m \left\{ C^* \left(\frac{k}{m} \tau \right) - C^* \left(\frac{k-1}{m} \tau \right) \right\} \frac{1}{\tau} f_\beta \left(\frac{t}{\tau}, k, m - k + 1 \right), \quad (2.8)$$

where $f_\beta \left(\frac{t}{\tau}, k, m - k + 1 \right)$ is the probability density function of a Beta distribution with parameters k and $m - k + 1$ valued at $\frac{t}{\tau}$.

Now, assume $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_m)$, with

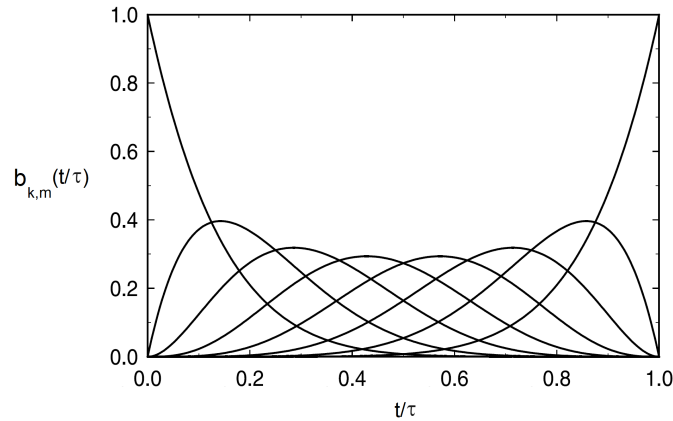
$$\varphi_k = C^* \left(\frac{k}{m} \tau \right) - C^* \left(\frac{k-1}{m} \tau \right); \varphi_k \geq 0, k = 1, \dots, m.$$

Note that $\boldsymbol{\varphi}$ is not time-dependent. Its values are unknown. Also consider $\mathbf{g}_m(t) = (g_{1,m}(t), \dots, g_{m,m}(t))'$, where

$$g_{k,m}(t) = \frac{1}{\tau} f_\beta \left(\frac{t}{\tau}, k, m - k + 1 \right); g_{k,m}(t) \geq 0, k = 1, \dots, m.$$

Thus, we can rewrite (2.8) as

$$b_m^{C^*}(t) = \boldsymbol{\varphi} \mathbf{g}_m(t).$$

Figure 2.2: The Bernstein basis functions of degree 7 on t in $[0, \tau]$.

Source: [Farouki \(2012\)](#).

[Osman and Ghosh \(2012\)](#) used this expression to model the hazard function. That is,

$$h(t|\boldsymbol{\varphi}) = \boldsymbol{\varphi} \mathbf{g}_m(t), t \in [0, \infty). \quad (2.9)$$

The authors justify the choice of Bernstein Polynomials for this purpose due to their attractive properties. More explicitly, they argue that since such polynomials have good derivation properties. Furthermore, the monotonicity of the cumulative hazard function is naturally modeled by Bernstein Polynomials, since that $\varphi_k \geq 0, \forall k \in \{1, \dots, m\}$. This function is expressed by

$$H(t|\boldsymbol{\varphi}) = \int_0^t h(u, \boldsymbol{\varphi}) du = \boldsymbol{\varphi} \mathbf{G}_m(t), \quad (2.10)$$

with

$$\mathbf{G}_m(t) = (G_{1,m}(t), \dots, G_{m,m}(t))',$$

where

$$G_{m,k}(t) = \int_0^t f_{\beta} \left(\frac{u}{\tau}; k, m - k + 1 \right) d \left(\frac{u}{\tau} \right), \forall k \in \{1, \dots, m\}.$$

The function $G_{m,k}(t)$ is the Beta cumulative distribution function with parameters k and $m - k + 1$.

[Osman and Ghosh \(2012\)](#) also discuss some aspects of choosing τ in the previous expressions. This must be done with care, as it influences the estimation. It is necessary that $\tau < \infty$, such that $\tau = \inf\{t : S(t) = 0\}$. In practice, in survival analysis, τ is chosen as the maximum value among the times observed until the occurrence of the event of interest or until the follow-up stops. Here, we will denote it by $\hat{\tau}$. But, using this choice, it is not possible to satisfy $H(\tau|\boldsymbol{\varphi}) = \infty$. Besides, there is no information about survival times in the region $t > \hat{\tau}$ ([Demarqui et al., 2019](#)). Therefore, this choice requires an adjustment in the hazard and cumulative hazard functions. As a solution, [Osman and](#)

Ghosh (2012) suggest some alterations in these functions, as follows:

$$h^*(t|\boldsymbol{\varphi}) = \begin{cases} h(t|\boldsymbol{\varphi}), & \text{if } 0 \leq t < \hat{\tau}, \\ m \frac{\varphi_m}{\hat{\tau}}, & \text{if } t \geq \hat{\tau}, \end{cases}$$

$$H^*(t|\boldsymbol{\varphi}) = \begin{cases} H(t|\boldsymbol{\varphi}), & \text{if } 0 \leq t < \hat{\tau}, \\ H(t|\boldsymbol{\varphi}) + m(t - \hat{\tau}) \frac{\varphi_m}{\hat{\tau}}, & \text{if } t \geq \hat{\tau}. \end{cases}$$

There are few works in the literature with applications of Bernstein Polynomials in survival analysis. One of them is Chang et al. (2005) which uses the Bernstein polynomials whose degree is a random quantity that needs to be estimated. Demarqui et al. (2019) use the Bernstein polynomials to model the baseline functions of the YP model. The R package YPBP of Demarqui (2020a) can be used to fit this model. Panaro (2020) developed an R package named `spsurv` (Panaro et al., 2020) to explain survival times using Bernstein polynomials coupled to some regression structures as PH and PO models.

2.5 Piecewise exponential model

The piecewise exponential model was introduced by Kalbfleisch and Prentice (1973) and is widely used in survival analysis due to its flexibility. It establishes a finite partition of the time axis and, for each partition, a constant hazard function is defined. Thus, the model allows approximating the hazard function using line segments.

The model is constructed as follows. Consider a time grid $\rho = \{\rho_0, \dots, \rho_m\}$. Thus, ρ makes a partition of the time axis in m intervals at the points $\rho_0, \rho_1, \dots, \rho_m$, with $0 = \rho_0 < \rho_1 < \dots < \rho_m < \infty$. The intervals generated from that partition are $I_1 = (\rho_0, \rho_1], I_2 = (\rho_1, \rho_2], \dots, I_m = (\rho_{m-1}, \rho_m]$.

The set ρ , and consequently the quantity of intervals m , can be established in different ways. Breslow (1974) and Demarqui and Mayrink (2021) assume that ρ is a known set composed by each of different time-to-event observations. The choice of ρ has influence over the inferential results, since we assume that the hazard function in each interval is constant and given by

$$h(t) = \lambda_j, \text{ for } t \in I_j, j = 1, \dots, m \text{ and } \lambda_j > 0. \quad (2.11)$$

Kalbfleisch and Prentice (1973) affirm that the choice of ρ can be independent of the data set. On the other hand, Demarqui and Mayrink (2021) argues that large m values can provide unstable estimates. In other approaches, ρ is treated as being random; see Demarqui et al. (2011, 2012).

Regardless of the choice of ρ , the cumulative hazard function is given by

$$H(t|\boldsymbol{\lambda}) = \sum_{j=1}^m \lambda_j (t_j - \rho_{j-1}), \quad (2.12)$$

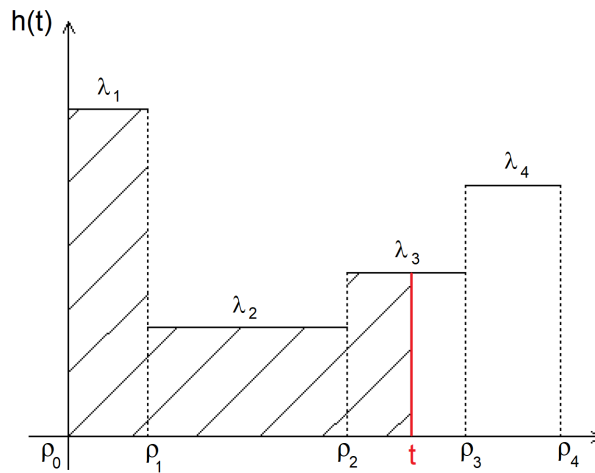
where

$$t_j = \begin{cases} \rho_{j-1}, & \text{if } t < \rho_{j-1}; \\ t, & \text{if } \rho_{j-1} < t \leq \rho_j; \\ \rho_j, & \text{if } t > \rho_j. \end{cases} \quad (2.13)$$

The piecewise exponential model takes the exponential model as a particular case when $m = 1$.

To understand the Expression (2.12), see the example illustrated by [de Mello \(2016\)](#) in Figure 2.3. In this case, it is assumed that the time axis has been divided into four intervals and that t is the time-to-event or censoring time. The cumulative hazard function can be interpreted as the area hatched in the aforementioned figure.

Figure 2.3: Example of $h(t)$ defined by PE.



Source: [de Mello \(2016\)](#).

2.6 Clayton copula

In survival analysis, there are situations in which more than one survival time is observed for the same individual. In these cases, the survival data are multivariate. It is reasonable to assume that there is an association among the multiple survival times

observed for the same individual. In particular, consider the case where an individual may experience successive recurrent events and a terminal event. The terminal event prevents new recurrent events from occurring. One possible approach to accommodate this association between survival times is through the use of copulas.

Clayton (1978) used copulas to treat bivariate survival data. Oakes (1989) showed that the Clayton copula is a specific case of the gamma-shared frailty model. Joe (2005) proved the asymptotic efficiency of the two-stage estimation process based on copula models. Li et al. (2019) developed a joint Bayesian frailty-copula model for situations in which individuals experience recurrent events and a terminal event.

Now, we discuss some mathematical aspects of copulas. Copulas are functions that connect univariate marginal distributions with their joint multivariate distribution (Nelsen, 2006). There are several copulas in the literature, each leading to a type of association between variables. Let's start by mathematically defining the copula function.

Definition 2.6.1. *A copula is a multivariate distribution whose marginals are uniform variables $U(0, 1)$. Consider $\mathbf{U} = (U_1, \dots, U_n) \in [0, 1]^n$. Let \bar{C} be a copula and θ be the copula association parameter (Nelsen, 2006). Thus,*

$$\bar{C}(u_1, \dots, u_n; \theta) = P(U_1 \leq u_1, \dots, U_n \leq u_n; \theta), \text{ with } (u_1, \dots, u_n) \in [0, 1]^n. \quad (2.14)$$

The existence of such copulas is guaranteed by Sklar's theorem.

Theorem 2.6.1. *(Sklar's theorem) Suppose that H is a joint distribution function and its margins are F_1, \dots, F_n . Then, for all $x_1, \dots, x_n \in \bar{\mathbb{R}}^n$, exists a copula \bar{C} such that*

$$H(x_1, \dots, x_n) = \bar{C}(F_1(x_1), \dots, F_n(x_n); \theta),$$

where $\bar{\mathbb{R}}^n = [-\infty, \infty]^n$. The function \bar{C} is unique if F_1, \dots, F_n are continuous; otherwise, \bar{C} is uniquely determined on $\text{Range}(F_1) \times \dots \times \text{Range}(F_n)$. Understand $\text{Range}(\cdot)$ as the image of a function. Conversely, consider that \bar{C} is a copula and F_1, \dots, F_n are distribution functions. Then, the function H is a joint distribution function and its margins are F_1, \dots, F_n (Hofert et al., 2018).

From Definition 2.6.1, we present the concept of the survival copula in Definition 2.6.2.

Definition 2.6.2. *Let $\mathbf{U} = (U_1, \dots, U_n) \in [0, 1]^n$ and θ be the copula association parameter. A survival copula is a function that connects marginal survival copulas with their joint distribution (Hofert et al., 2018). Mathematically,*

$$C(u_1, \dots, u_n; \theta) = P(U_1 > u_1, \dots, U_n > u_n; \theta), \text{ with } (u_1, \dots, u_n) \in [0, 1]^n. \quad (2.15)$$

From on now, we will focus on bivariate survival copulas, that is, when $n = 2$ in Definition 2.6.2. This is because, in this study, we will use survival copulas to accommodate the correlation between just two types of events.

Nelsen (2006) highlights a very important result about the existence of first-order partial derivatives of copulas, which will be useful when dealing with our likelihood function. This result is presented in Theorem 2.6.2.

Theorem 2.6.2. (Nelsen, 2006) *Let C be a bivariate copula. Then, for almost all u , and for such u and v ,*

$$\exists \frac{\partial}{\partial u} C(u, v; \theta) \text{ and } 0 \leq \frac{\partial}{\partial u} C(u, v; \theta) \leq 1.$$

Additionally, for almost all v , and for such u and v ,

$$\exists \frac{\partial}{\partial v} C(u, v; \theta) \text{ and } 0 \leq \frac{\partial}{\partial v} C(u, v; \theta) \leq 1.$$

Hofert et al. (2018) present a result about the copula's second-order derivatives $\partial^2 C / \partial u \partial v$. Through the chain rule, we can operate this derivation whose result is

$$f_{UV}(u, v) = f_U(u) f_V(v) \frac{\partial^2}{\partial u \partial v} C(u, v; \theta), \quad (2.16)$$

where $f_U(u)$ and $f_V(v)$ are the univariate marginal densities. The function $f_{UV}(u, v)$ is a bivariate density function, and this result will be useful when building the likelihood function.

To indicate the mentioned derivatives, consider the following notation

$$C_{(01)}(u, v; \theta) = \frac{\partial}{\partial v} C(u, v; \theta),$$

$$C_{(10)}(u, v; \theta) = \frac{\partial}{\partial u} C(u, v; \theta),$$

and

$$C_{(11)}(u, v; \theta) = \frac{\partial^2}{\partial u \partial v} C(u, v; \theta).$$

A great advantage of copulas is the clearer definition of a measure of the correlation between variables. This motivated the choice of the approach of copulas in the present work. Although copula already has an association parameter, researchers are usually interested in some measure of correlation. The most common is the Kendall's tau.

Definition 2.6.3. *Consider that (U_1, V_1) and (U_2, V_2) are two independent replicas of any pair of any random variables (U, V) . Kendall's tau τ_κ coefficient is defined as*

$$\tau_\kappa = P[(U_1 - U_2)(V_1 - V_2) > 0] - P[(U_1 - U_2)(V_1 - V_2) < 0].$$

Kendall's τ_κ measures the difference between the probabilities of concordance and discordance. Understand concordance as the cases where $U_1 > U_2$ and $V_1 > V_2$, or $U_1 < U_2$ and $V_1 < V_2$ and discordance, otherwise. In other words, we can say that when the value of one variable increases and the other also increases (alternatively, when the value of one variable decreases and of the other also decreases), the probability of concordance also increases (Nelsen, 2006).

If U and V are continuous, τ_κ can be calculated using

$$\tau_\kappa = 4 \int_0^1 \int_0^1 C(u, v; \theta) dC(u, v; \theta) - 1. \quad (2.17)$$

Kendall's τ_κ is one of the best-known correlation measures used in the context of copulas (Hofert et al., 2018) because it is invariant over monotonous transformations in opposite to Pearson's correlation coefficient. In this section, we start presenting some concepts about the Archimedean copula class, focusing on the bivariate case.

Definition 2.6.4. (Hofert et al., 2018) *A bivariate Archimedean copula has the form*

$$C(u, v; \theta) = \Upsilon (\Upsilon^{-1}(u) + \Upsilon^{-1}(v)), \quad (2.18)$$

where Υ is called the copula generating function and is continuous and descending on $[0, \infty]$ satisfying:

1. $\Upsilon(0) = 1$
2. $\Upsilon(\infty) = \lim_{t \rightarrow \infty} \Upsilon(t) = 0$, and
3. Υ is strictly decreasing in $[0, \inf\{t : \Upsilon(t) = 0\}]$.

Definition 2.6.4 shows the main difference between this class of copulas and the others, which is the fact that it is possible to write the copula function from a generator function. This brings some advantages because it allows one to generate copulas of this class by changing only the generating function.

Nelsen (2006) discusses some properties of bivariate Archimedean copulas:

1. Symmetry: $C(u, v; \theta) = C(v, u; \theta)$;
2. C is associative: $C(C(u, v; \theta), w; \theta) = C(u, C(v, w; \theta); \theta)$ for all $u, v, w \in [0, 1]$;
3. Let Υ be the generator of C . Then, for any constant $a > 0$, $a\Upsilon$ is also a generating function.

The Clayton Archimedean copula is considered in this work. Assume the marginal survival functions $u = S_1(t_1)$ and $v = S_2(t_2)$. Then, the joint survival function is

$$S_{T_1, T_2}(t_1, t_2) = C(u, v; \theta).$$

The Clayton copula is an Archimedean copula whose generating function is

$$\Upsilon(t) = (1 + t)^{-1/\theta}, \text{ with } \theta \in (0, \infty).$$

Then,

$$C(u, v; \theta) = (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}. \quad (2.19)$$

When $\theta \rightarrow 0$, U and V are independent because

$$\lim_{\theta \rightarrow 0} C(u, v; \theta) = uv,$$

and when $\theta \rightarrow \infty$, U and V have perfect positive dependency (Hofert et al., 2018). The Kendall's tau τ_κ for Clayton copula can be calculated using Expression (2.17) and its value is

$$\tau_\kappa = \frac{\theta}{2 + \theta},$$

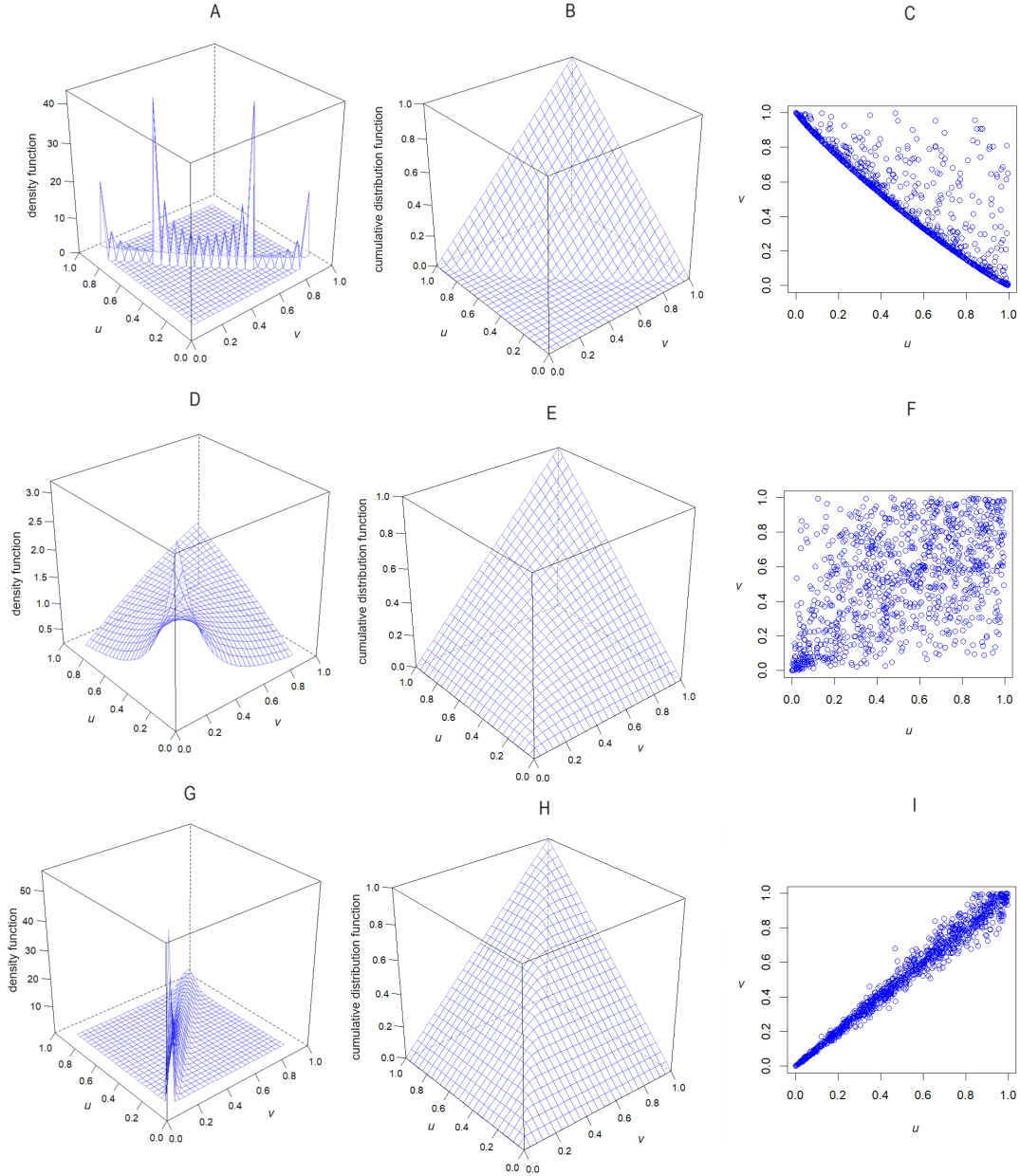
with $\tau_\kappa > 0$. There is an extended form of the Clayton copula in which $\theta \in [-1, \infty) - \{0\}$. In that case, $\tau_\kappa \in [-1, 1) - 0$. The purpose of this extension is to allow negative associations between the variables. It is defined by

$$C(u, v; \theta) = \max \{ (u^{-\theta} + v^{-\theta} - 1), 1 \}^{-\frac{1}{\theta}}.$$

To explore the graphical characteristics of this copula, refer to Figure 2.4, which presents plots generated using three distinct values of θ . For $\theta = -0.9$ ($\tau_\kappa \approx -0.82$), the density function of the Clayton copula is depicted in Figure 2.4-(A), with the corresponding distribution function shown in Figure 2.4-(B). Additionally, 1000 independent pairs of observations, derived from the Clayton copula with this θ value, are displayed in Figure 2.4-(C). This analysis was repeated for $\theta = 1$ ($\tau_\kappa \approx 0.33$), with the copula density function shown in Figure 2.4-(D), its distribution function in Figure 2.4-(E), and the 1000 independent observation pairs in Figure 2.4-(F). Finally, for $\theta = 20$ ($\tau_\kappa \approx 0.91$), the copula density function is presented in Figure 2.4-(G), the distribution function in Figure 2.4-(H), and the 1000 observation pairs in Figure 2.4-(I). For $\theta = -0.9$, the density function exhibits a hyperbolic pattern where high values of one variable tend to correspond with low values of the other. This pattern is mirrored in both the distribution function and the scatter plot. For $\theta = 1$, the density function shows a concentration along the line $u_2 = u_1$. When comparing the distributions for $\theta = 1$ and $\theta = 20$, it is evident that they differ; the surface generated with $\theta = 1$ is smoother. With $\theta = 20$, the scatter plot displays a very tight clustering of points along the line $u_2 = u_1$, indicating strong positive dependence, as high values of one variable almost invariably align with high values of the other. Conversely, there is more dispersion in the scatter plot when $\theta = 1$.

After the discussion on the graphical characteristics of the Clayton copula (Figure 2.4) we now focus on its mathematical formulation. Specifically, the first and second-order

Figure 2.4: Clayton copula, when $\theta = -0.9$: (A) density function, (B) cumulative distribution function, and (C) scatter plot. Clayton copula, when $\theta = 1$: (D) density function, (E) cumulative distribution function, and (F) scatter plot. Clayton copula, when $\theta = 20$: (G) density function, (H) cumulative distribution function, and (I) scatter plot.



Source: Prepared by the author.

derivatives of the Clayton copula function are detailed below:

$$C_{(01)}(u, v; \theta) = v^{-(\theta+1)}(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}-1},$$

$$C_{(10)}(u, v; \theta) = u^{-(\theta+1)}(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}-1},$$

and

$$C_{(11)}(u, v; \theta) = (\theta + 1)(uv)^{-(\theta+1)}(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}-2}.$$

In this thesis, we chose to use the simplest version of the Clayton copula, such that $\theta \in (0, \infty)$.

Chapter 3

Proposed models

In this chapter, we provide a discussion of the structure of our model classes. Section 3.1 focuses specifically on the Yang and Prentice models, addressing both individual and shared frailties. Within this section, we define the notation employed in the models, outline the types of data that are suitable for modeling, and detail the likelihood function used to estimate the parameters of interest. Section 3.2 explores the particularities of our second class of models, which is focused on data involving recurrent and terminal events. In this part of the text, we introduce the joint frailty-copula models and explain the notation employed, followed by a characterization of the regression families and their likelihood function. The parameter estimation process is done applying the `rstan` package (Stan Development Team, 2018).

3.1 Class 1: Yang and Prentice frailty model

In certain situations, survival data can exhibit correlations due to natural or artificial groupings among individuals (Colosimo and Giolo, 2006; Li et al., 2019). For example, survival times may be observed within the same family or among patients treated in the same intensive care unit (ICU). It is reasonable to assume that individuals within the same cluster share some similarities in their survival times. Therefore, considering independence among survival times within a group may not be a realistic assumption. In such cases, a common approach is to introduce a shared frailty term in the hazard function, where all individuals within a group share the same effect.

Alternatively, survival data may involve individuals experiencing recurrent events, where the survival times of the same individual exhibit a correlation with each other. For instance, patients who experience recurrent strokes while hospitalized. In this scenario, individuals can be considered as a group, and a random effect can be introduced for each individual, with the assumption that all survival times within the same individual are independent given the frailty.

Frailty can also be employed in the context of univariate survival data, where each individual is considered a unit-sized group. In this context, frailty is incorporated to account for unobserved heterogeneity and explain why individuals with similar observed characteristics may exhibit distinct survival times (Vaupel et al., 1979; Colosimo and Giolo, 2006).

In the cases described above, the key distinction lies in how the group is defined, within which survival data are conditionally independent given the frailty. This section commences with an introduction to the notation and the likelihood function for clustered data in subsection 3.1.1. Following this, subsection 3.1.2 presents the notation and the likelihood function for data with recurrent events.

One goal of this study is to propose a Bayesian frailty model in a Yang and Prentice regression structure. In this approach, the baseline hazard functions are modeled using the exponential function, the piecewise exponential model, and Bernstein polynomials. The modeling of $h_0(t)$, the baseline hazard function, is achieved by adopting (2.9) (Demarqui et al., 2019; Panaro, 2020) for the Bernstein polynomials model, and (2.11) (Breslow, 1972, 1974; Schneider et al., 2020) for the piecewise exponential model.

3.1.1 Notation and the likelihood function for clustered data

Consider a study with L clusters, where the group sizes are denoted as n_1, n_2, \dots, n_L , which may vary across groups. Let $R_{i,j}$ be the time-to-event of the individual j ; $j = 1, \dots, n_L$, of cluster i , with $i = 1, \dots, L$. Suppose that the survival times $R_{i,j}, \dots, R_{L,n_L}$ are mutually independent conditioned on the frailty of the cluster w_i . Define $C_{i,j}$ as the time until administrative censoring and $\delta_{i,j} = I(R_{i,j} < C_{i,j})$, the failure state indicator where $\delta_{i,j} = 0$ means administrative censoring and $\delta_{i,j} = 1$, an event. Let $Y_{i,j} = \min\{R_{i,j}, C_{i,j}\}$ be the observed time of the j -th individual of the i -th cluster. Figure 3.1 illustrates the notation used in this section. Two possible situations can be evaluated according to the notation established.

- The first is when each time-to-event concerns a unique individual and the clusters represent groups of individuals who share the element of frailty. That is, $n_i \geq 1, \forall i \in \{1, \dots, L\}$.
- The second case involves databases in which each individual has only one survival time, therefore each cluster is a one-point set. That is, $n_i = 1, \forall i \in \{1, \dots, L\}$.

To model the times to events, we employ the YP family. The survival function, in

this case, can be expressed as:

$$S(r_{i,j}|\mathbf{x}_{i,j}, w_i) = \left[1 + \frac{\nu_{i,j}}{\xi_{i,j}} \mathcal{R}_0(r_{i,j}) \right]^{\xi_{i,j}}. \quad (3.1)$$

where

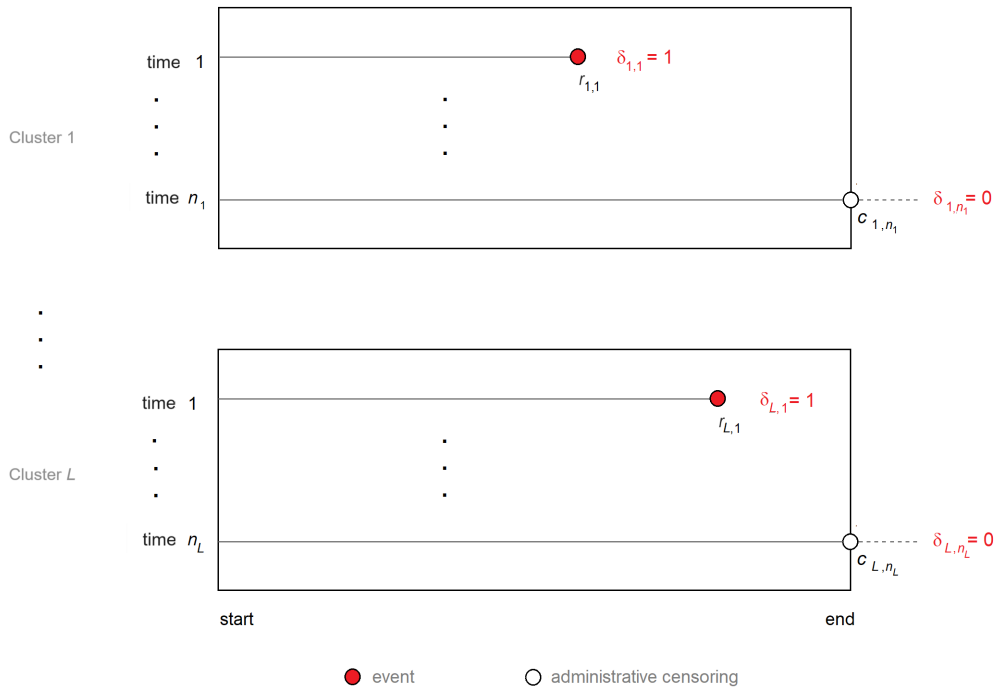
$$\nu_{i,j} = \exp(\mathbf{x}_{i,j}\boldsymbol{\psi} + w_i)$$

and

$$\xi_{i,j} = \exp(\mathbf{x}_{i,j}\boldsymbol{\phi} + w_i).$$

The term $\nu_{i,j}$ represents the short-term hazard ratio, $\xi_{i,j}$ represents the long-term hazard ratio in the YP model, $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ are $p \times 1$ vectors of regression coefficients, and \mathbf{x}_i is a $1 \times p$ vector of covariates of the regression. It is important to note that incorporating frailty into the YP model is a contribution of this work.

Figure 3.1: Schematic representation of clustered survival times.



Source: Prepared by the author.

We can write the likelihood function as:

$$\mathcal{L}(\Theta|\mathcal{D}, \mathbf{w}) = \prod_{i=1}^L \prod_{j=1}^{n_i} [f(y_{i,j}|w_i)]^{\delta_{i,j}} [S(y_{i,j}|w_i)]^{1-\delta_{i,j}},$$

where \mathcal{D} is the set of observed data, such that

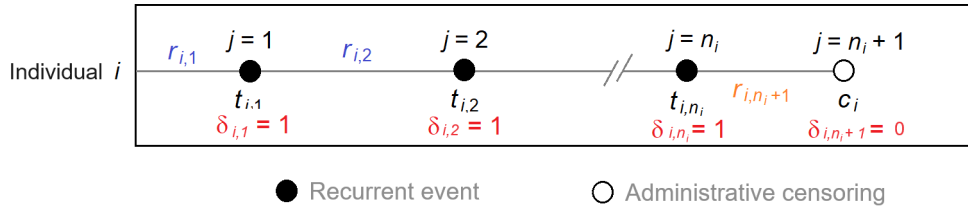
$$\mathcal{D} = \{y_{i,j}, \delta_{i,j}, \mathbf{x}_{i,j}; i = 1, \dots, L; j = 1, \dots, n_i\},$$

and $\Theta = \{\boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\phi}, \sigma_w^2\}$ denotes the set of parameters to be estimated in the models.

3.1.2 Notation and the likelihood function for data with recurrent events

Now, consider L to be the number of individuals. Denote by C_i the time to the administrative censoring, that is, the time until loss of follow-up for some reason external to the study, with $i = 1, \dots, L$. Denote by $R_{i,j}$ the gap-time between the $(j-1)$ -th and j -th occurrences of the recurrent event, and let $T_{i,j} = \sum_{j'=1}^j R_{i,j'}$ be the total observation time until the j -th recurrent event. Suppose the i -th individual experiences a total of n_i recurrent events. When $j = n_i + 1$, $R_{i,n_i+1} = C_i - \sum_{j=1}^{n_i} R_{i,j}$, which can be interpreted as the gap-time between the n_i -th recurrent event and the end of follow-up. Define $\delta_{i,j} = I(T_{i,j} < C_i)$ is the failure state indicator for the j -th recurrent event. When $\delta_{i,j} = 0$, it indicates that the observed time is an administrative censoring time. The notation used in this section is illustrated by Figure 3.2

Figure 3.2: Schematic representation of data with recurrent events.



Source: Prepared by the author.

We employ the YP family to model the times to events whose survival function is also expressed by (3.1). Assuming that the survival times $R_{i,j}, \dots, R_{L,n_L}$ are mutually independent conditioned on the frailty term w_i , we can obtain the likelihood function as:

$$\mathcal{L}(\Theta|\mathcal{D}, \mathbf{w}) = \prod_{i=1}^L \left\{ S(r_{i,n_i+1}|w_i) \prod_{j=1}^{n_i} f(r_{i,j}|w_i) \right\},$$

where \mathcal{D} as the set of observed data, such that

$$\mathcal{D} = \{r_{i,j}, r_{i,n_i+1}, \delta_{i,j}, \mathbf{x}_{i,j}; i = 1, \dots, L; j = 1, \dots, n_i\}.$$

Let $\Theta = \{\gamma, \psi, \phi, \sigma_w^2\}$ denote the set of parameters to be estimated in the models.

3.2 Class 2: The joint frailty-copula models

In this section, we discuss the joint frailty-copula model for multivariate survival data. The data analyzed here consists of the survival times of individuals who may experience both recurrent and terminal events. We introduce the notation in subsection 3.2.1 and present the likelihood along with its construction steps in subsection 3.2.2. This section also highlights the proposed model class and its contributions to the literature.

3.2.1 Notation

Let L be the number of individuals. Each individual can experience two types of events: recurrent and terminal events. Denote by D_i the time to the terminal event and by C_i the time to the administrative censoring, that is, the time until loss of follow-up for some reason external to the study. Let $\delta_i = I(D_i < C_i)$ be the failure state indicator for the terminal event. The observable time of i -th individual is then given by $Y_i = \min\{D_i, C_i\}$.

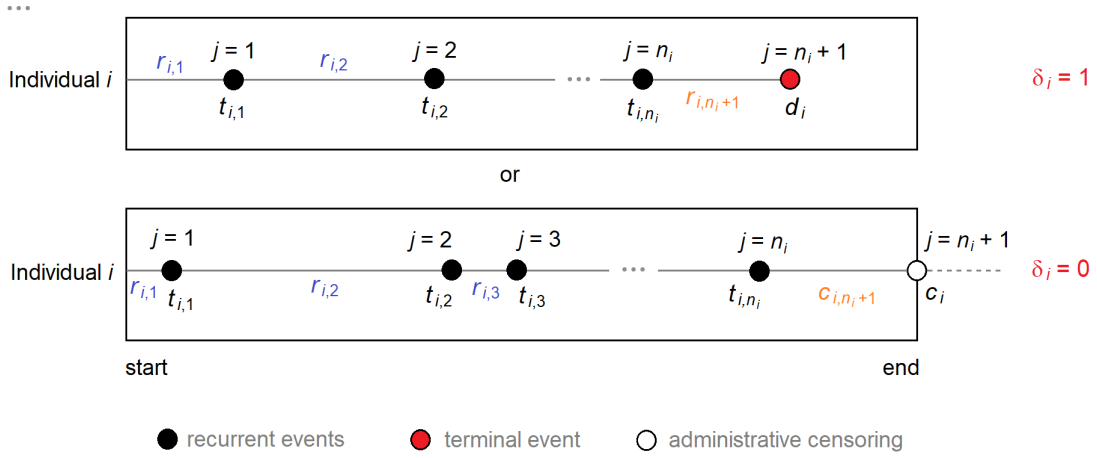
Now, denote by $R_{i,j}$ the gap-time between the $(j-1)$ -th and j -th occurrences of the recurrent event, and let $T_{i,j} = \sum_{j'=1}^j R_{i,j'}$ be the total observation time until the j -th recurrent event. Suppose the i -th subject experiences a total of n_i recurrent events. When $j = n_i + 1$, $R_{i,n_i+1} = Y_i - \sum_{j=1}^{n_i} R_{i,j}$, which can be interpreted as the gap-time between the n_i -th recurrent event and the end of follow-up.

When the terminal event occurs, it generates an informative censoring on the recurrent event process, once it prevents new recurrent events from happening to an individual. Define $C_{i,j} = \max\{Y_i - T_{i,j-1}, 0\}$ as the time to this dependent censoring on the j -th recurrent event. A schematic view of the notation adopted is seen in Figure 3.3.

An approach based on survival copulas will be used to accommodate the correlation between recurrent events and the terminal event. Three families of survival regression models will be used to explain the effect of some covariates on the time to both types of events: PH, PO, and YP families. In all of them, consider $\mathbf{x}_i^{(T)}$ and $\mathbf{x}_i^{(R)}$ as $1 \times p$ vectors of covariates of the regression model associated with the terminal event and recurrent events, respectively, and w_i as the individual frailty element which captures the correlation between its recurrent events. It is assumed that the frailty element w_i has a distribution $N(0, \sigma_w^2)$.

In the context of the PH and PO families, denote $\boldsymbol{\beta}^{(T)}$ and $\boldsymbol{\beta}^{(R)}$ as $p \times 1$ vectors. These vectors quantify the impacts of the covariates $\mathbf{x}_i^{(T)}$ and $\mathbf{x}_i^{(R)}$ on D_i and $R_{i,j}$, respec-

Figure 3.3: Schematic representation of multivariate survival times.



Source: Prepared by the author.

tively. Specifically, $\boldsymbol{\beta}^{(T)}$ represents the regression coefficients associated with the terminal event, while $\boldsymbol{\beta}^{(R)}$ corresponds to those for recurrent events. In the case of YP families, we introduce $\boldsymbol{\psi}^{(T)}$ and $\boldsymbol{\phi}^{(T)}$, vectors quantifying the short-term and long-term effects of the covariates $\mathbf{x}_i^{(T)}$ on terminal events, respectively. Similarly, $\boldsymbol{\psi}^{(R)}$ and $\boldsymbol{\phi}^{(R)}$ are vectors that measure the short-term and long-term effects of the covariates $\mathbf{x}_i^{(T)}$ on recurrent events.

For the PH family, the hazard function for the time to terminal event D_i , and the gap-time to recurrent event $R_{i,j}$ of the i -th subject are expressed by

$$h_D \left(d | \mathbf{x}_i^{(T)}, w_i \right) = h_0^{(T)}(d) \exp \left(\mathbf{x}_i^{(T)} \boldsymbol{\beta}^{(T)} + w_i \right) \quad (3.2)$$

and

$$h_R \left(r | \mathbf{x}_i^{(R)}, w_i \right) = h_0^{(R)}(r) \exp \left(\mathbf{x}_i^{(R)} \boldsymbol{\beta}^{(R)} + w_i \right). \quad (3.3)$$

The baseline hazard functions $h_0^{(T)}(\cdot)$ and $h_0^{(R)}(\cdot)$ will be modeled using the exponential, Bernstein polynomials, and piecewise exponential models.

The PO family, discussed in Section 2.2.2, will also be applied to model the times of both terminal and recurrent events. The respective odds functions, in this case, will be given by

$$\mathcal{R}_D \left(d | \mathbf{x}_i^{(T)}, w_i \right) = \mathcal{R}_0^{(T)}(d) \exp \left(\mathbf{x}_i^{(T)} \boldsymbol{\beta}^{(T)} + w_i \right)$$

and

$$\mathcal{R}_R \left(r | \mathbf{x}_i^{(R)}, w_i \right) = \mathcal{R}_0^{(R)}(r) \exp \left(\mathbf{x}_i^{(R)} \boldsymbol{\beta}^{(R)} + w_i \right). \quad (3.4)$$

Finally, the YP family will also be employed to model the times to the terminal and recurrent events. In this case, we will have the following survival functions:

$$S_D \left(d | \mathbf{x}_i^{(T)}, w_i \right) = \left[1 + \frac{\nu^{(T)}}{\xi^{(T)}} \mathcal{R}_0(d) \right]^{\xi^{(T)}} \quad (3.5)$$

and

$$S_R \left(r | \mathbf{x}_i^{(R)}, w_i \right) = \left[1 + \frac{\nu^{(R)}}{\xi^{(R)}} \mathcal{R}_0(r) \right]^{\xi^{(R)}}. \quad (3.6)$$

Here,

$$\begin{aligned} \nu^{(T)} &= \exp \left(\mathbf{x}_i^{(T)} \boldsymbol{\psi}^{(T)} + w_i \right), \quad \xi^{(T)} = \exp \left(\mathbf{x}_i^{(T)} \boldsymbol{\phi}^{(T)} + w_i \right), \\ \nu^{(R)} &= \exp \left(\mathbf{x}_i^{(R)} \boldsymbol{\psi}^{(R)} + w_i \right), \quad \text{and } \xi^{(R)} = \exp \left(\mathbf{x}_i^{(R)} \boldsymbol{\phi}^{(R)} + w_i \right), \end{aligned}$$

where $\nu^{(T)}$ and $\nu^{(R)}$ are short-term hazard ratios and $\xi^{(T)}$ and $\xi^{(R)}$ are long-term hazard ratios in YP model. The incorporation of frailty in the YP model constitutes a contribution from the present work.

3.2.2 The likelihood function

Consider that n_i recurrent events occur for the individual i . According to [Li et al. \(2019\)](#), the joint probability of $D_i > d_i$ and $R_{i,j} > r_{i,j}$, conditional on w_i , is

$$P(D_i > d_i, R_{i,j} > r_{i,j} | w_i) = C[S_D(d_i | w_i), S_R(r_{i,j} | w_i)], \forall i = 1, \dots, L; \forall j = 1, \dots, n_i + 1,$$

where C is the copula function as defined in Chapter 2. Based on the survival copula, we have

$$\begin{aligned} P(D_i > d_i, R_{i,j} = r_{i,j} | w_i) &= C_{(01)}[S_D(d_i | w_i), S_R(r_{i,j} | w_i); \theta] f_R(r_{i,j} | w_i), \\ P(D_i = d_i, R_{i,j} = r_{i,j} | w_i) &= C_{(11)}[S_D(d_i | w_i), S_R(r_{i,j} | w_i); \theta] f_D(d_i | w_i) f_R(r_{i,j} | w_i) \end{aligned}$$

and

$$P(D_i = d_i, R_{i,j} > r_{i,j} | w_i) = C_{(10)}[S_D(d_i | w_i), S_R(r_{i,j} | w_i); \theta] f_D(d_i | w_i),$$

where $C_{(01)} = \frac{\partial}{\partial v} C(u, v; \theta)$, $C_{(10)} = \frac{\partial}{\partial u} C(u, v; \theta)$ and $C_{(11)} = \frac{\partial^2}{\partial u \partial v} C(u, v; \theta)$. For a subject with $\delta_i = 0$, $C_{i,j} = \max\{C_i - T_{i,j-1}, 0\}$ and $C_{i,j}$ is independent of $R_{i,j}$. The aforementioned probabilities represent contributions of the gap times between recurrences and the times until the terminal event on the likelihood function.

When $\delta_i = 0$, the individual i has not experienced the terminal event, because it was censored. Given the assumption that $R_{i,j}, \dots, R_{i,n_i+1}$ are mutually independent, conditional on w_i , the probability of the i -th subject survives up to d_i and experiences n_i

recurrent events is (Li et al., 2019)

$$\begin{aligned}
& P(D_i > d_i, R_{i,1} = r_{i,1}, \dots, R_{i,n_i} = r_{i,n_i}, R_{i,n_i+1} > c_{i,n_i+1} | w_i) = \\
& = P(R_{i,1} = r_{i,1}, \dots, R_{i,n_i} = r_{i,n_i}, R_{i,n_i+1} > c_{i,n_i+1} | D_i > d_i, w_i) P(D_i > d_i | w_i) \\
& = P(R_{i,1} = r_{i,1}, \dots, R_{i,n_i} = r_{i,n_i} | D_i > d_i, w_i) P(R_{i,n_i+1} > c_{i,n_i+1} | D_i > d_i, w_i) \\
& \quad \times P(D_i > d_i | w_i) \\
& = \prod_{j=1}^{n_i} P(R_{i,j} = r_{i,j} | D_i > d_i, w_i) P(R_{i,n_i+1} > c_{i,n_i+1} | D_i > d_i, w_i) P(D_i > d_i | w_i) \\
& = \prod_{j=1}^{n_i} \frac{P(R_{i,j} = r_{i,j}, D_i > d_i | w_i)}{P(D_i > d_i | w_i)} \frac{P(R_{i,n_i+1} > c_{i,n_i+1}, D_i > d_i | w_i)}{P(D_i > d_i | w_i)} P(D_i > d_i | w_i) \\
& = \prod_{j=1}^{n_i} \frac{P(R_{i,j} = r_{i,j}, D_i > d_i | w_i)}{S_D(d_i | w_i)} \frac{P(R_{i,n_i+1} > c_{i,n_i+1}, D_i > d_i | w_i)}{S_D(d_i | w_i)} S_D(d_i | w_i) \\
& = \prod_{j=1}^{n_i} \frac{C_{(01)} [S_D(d_i | w_i), S_R(r_{i,j} | w_i); \theta] f_R(r_{i,j} | w_i)}{S_D(d_i | w_i)} C [S_D(d_i | w_i), S_R(c_{i,n_i+1}; \theta | w_i)].
\end{aligned} \tag{3.7}$$

On the other hand, if $\delta_i = 1$, the individual i has experienced the terminal event. Thus, given the same assumption that $R_{i,j}, \dots, R_{i,n_i+1}$ are mutually independent, conditional on w_i , we can express the probability that the i -th subject survives until d_i and experiences n_i events as given below (Li et al., 2019):

$$\begin{aligned}
& P(D_i = d_i, R_{i,1} = r_{i,1}, \dots, R_{i,n_i} = r_{i,n_i}, R_{i,n_i+1} > c_{i,n_i+1} | w_i) = \\
& = P(R_{i,1} = r_{i,1}, \dots, R_{i,n_i} = r_{i,n_i} | D_i = d_i, w_i) P(R_{i,n_i+1} > c_{i,n_i+1} | D_i = d_i, w_i) \\
& \quad \times P(D_i = d_i | w_i) \\
& = \prod_{j=1}^{n_i} P(R_{i,j} = r_{i,j} | D_i = d_i, w_i) P(R_{i,n_i+1} > c_{i,n_i+1} | D_i = d_i, w_i) P(D_i = d_i | w_i) \\
& = \prod_{j=1}^{n_i} \frac{P(R_{i,j} = r_{i,j}, D_i = d_i | w_i)}{P(D_i = d_i | w_i)} \frac{P(R_{i,n_i+1} > c_{i,n_i+1}, D_i = d_i | w_i)}{P(D_i = d_i | w_i)} P(D_i = d_i | w_i) \\
& = \prod_{j=1}^{n_i} \frac{P(R_{i,j} = r_{i,j}, D_i = d_i | w_i)}{f_D(d_i | w_i)} \frac{P(R_{i,n_i+1} > c_{i,n_i+1}, D_i = d_i | w_i)}{f_D(d_i | w_i)} f_D(d_i | w_i) \\
& = \prod_{j=1}^{n_i} C_{(11)} [S_D(d_i | w_i), S_R(r_{i,j} | w_i); \theta] f_R(r_{i,j} | w_i) \\
& \quad \times C_{(10)} [S_D(d_i | w_i), S_R(c_{i,n_i+1} | w_i); \theta].
\end{aligned} \tag{3.8}$$

Define \mathcal{D}_L as the set of observed data, such that

$$\mathcal{D}_L = \left\{ y_i, \delta_i, \delta_{i,j}, r_{i,j}, \mathbf{x}_i^{(T)}, \mathbf{x}_i^{(R)}; i = 1, \dots, L; j = 1, \dots, n_i + 1 \right\}.$$

Denote by $\beta^* = \{\beta^{(T)}, \beta^{(R)}\}$ the set of regression coefficients in PH and PO families and $\psi^* = \{\psi^{(T)}, \psi^{(R)}, \phi^{(T)}, \phi^{(R)}\}$ the regression coefficients in YP family. In addition, assume

$\Theta = \{\boldsymbol{\gamma}^{(T)}, \boldsymbol{\gamma}^{(R)}, \boldsymbol{\beta}^*, \sigma_w^2, \theta\}$ the set of parameters of models to be estimated, where $\boldsymbol{\gamma}^{(T)}$ and $\boldsymbol{\gamma}^{(R)}$ are the parameters of the baseline hazard function. Thus, the conditional likelihood function is given by

$$\begin{aligned} \mathcal{L}(\Theta|\mathcal{D}_L, \mathbf{w}) &= \prod_{i=1}^L \{P(D_i > y_i, R_{i,1} = r_{i,1}, \dots, R_{i,n_i} = r_{i,n_i}, R_{i,n_i+1} > r_{i,n_i+1}|w_i)\}^{1-\delta_i} \\ &\quad \times \{P(D_i = y_i, R_{i,1} = r_{i,1}, \dots, R_{i,n_i} = r_{i,n_i}, R_{i,n_i+1} > r_{i,n_i+1}|w_i)\}^{\delta_i}. \end{aligned}$$

Using the results (3.7) and (3.8), we have the likelihood function (Li et al., 2019):

$$\begin{aligned} \mathcal{L}(\Theta|\mathcal{D}_L, \mathbf{w}) &= \\ &= \prod_{i=1}^L \left\{ \prod_{j=1}^{n_i} \frac{C_{(01)}[S_D(y_i|w_i), S_R(r_{i,j}|w_i)] f_R(r_{i,j}|w_i)}{S_D(y_i|w_i)} C[S_D(y_i|w_i), S_R(r_{i,n_i+1}|w_i)] \right\}^{1-\delta_i} \\ &\quad \times \left\{ \prod_{j=1}^{n_i} \frac{C_{(11)}[S_D(y_i|w_i), S_R(r_{i,j}|w_i)] f_D(y_i|w_i) f_R(r_{i,j}|w_i)}{f_D(y_i|w_i)} \right\}^{\delta_i} \\ &\quad \times \left\{ \prod_{j=1}^{n_i} C_{(10)}[S_D(y_i|w_i), S_R(r_{i,n_i+1}|w_i)] \right\}^{\delta_i} \\ &= \prod_{i=1}^L \left\{ \frac{C_{(10)}[S_D(y_i|w_i), S_R(r_{i,n_i+1}|w_i)]}{f_D(y_i|w_i)} \right\}^{\delta_i} \times \left\{ \frac{C[S_D(y_i|w_i), S_R(r_{i,n_i+1}|w_i)]}{[S_D(y_i|w_i)]^{n_i}} \right\}^{1-\delta_i} \\ &\quad \times \prod_{j=1}^{n_i} \left\{ C_{(01)}[S_D(y_i|w_i), S_R(r_{i,j}|w_i)] f_R(r_{i,j}|w_i) \right\}^{1-\delta_i} \\ &\quad \times \left\{ C_{(11)}[S_D(y_i|w_i), S_R(r_{i,j}|w_i)] f_D(y_i|w_i) f_R(r_{i,j}|w_i) \right\}^{\delta_i}. \end{aligned} \tag{3.9}$$

The present work also proposes an extension of the joint frailty-copula model of Li et al. (2019) in a Bayesian approach and this extension is one of our contributions to the literature. The authors propose a model that utilizes only the PH regression structure and exponential baseline functions, incorporating the Clayton copula for capturing dependence. The objective here is to expand that model by incorporating two additional regression structures, PO and YP, in which baseline functions will be modeled by the Bernstein polynomials and the piecewise exponential. The application of the Bernstein polynomials and the piecewise exponential model to fit the baseline hazard functions is done by assuming $h_0(t)$ as described in (2.9) (Demarqui et al., 2019; Panaro, 2020) and in (2.11) (Breslow, 1972, 1974; Schneider et al., 2020), respectively.

Chapter 4

Monte Carlo simulation study

In this chapter, we detail our simulation studies designed to assess the impact of model selection on parameter estimates. Our goal is to investigate the estimation biases, average standard error, posterior standard deviation, credible intervals, and coverage probability. The simulation study was conducted, with data generation and model fitting executed in the R programming language (R Core Team, 2024). We utilized `rstan` package (Stan Development Team, 2018) to generate four Markov chain Monte Carlo (MCMC) chains for each parameter, each chain comprising 2000 iterations, with 1000 warm-up iterations. This approach yielded posterior sample sizes of 4000 for each parameter.

Our studies were conducted for both classes of models. The first class comprises the YP frailty models, in which we evaluated two scenarios:

- individual frailty, where each individual represents a single-unit cluster, and
- shared frailty, where individuals are clusters of size greater than or equal to one. In this case, we mean that individuals experience recurrent events.

The detailed steps of the simulation, its specific configurations, and the outcomes are discussed in Section 4.1. Additionally, we run another simulation study for the second class of models, which encompasses the joint frailty-copula models. An overview of the settings and steps undertaken for this simulation, along with an analysis of the results are presented in Section 4.2.

We now proceed with the details of generating survival times. Given that the YP model generalizes the PH and PO regression families, we have chosen to generate our data using this model. Consider an individual with p characteristics denoted by $\mathbf{x} = (x_1, \dots, x_p)$. For our simulation studies, we define the specific values for the short-term regression coefficients $\boldsymbol{\psi} = (\psi_1, \dots, \psi_p)'$ and the long-term coefficients $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$. Additionally, we established a determined value for the variance of the frailty term, denoted as σ_w^2 . We assume that the frailty w follows a normal distribution with mean 0 and variance σ_w^2 . We calculate linear predictors of our models η_S and η_L . The short-term and long-term linear predictors for an individual are calculated using $\eta_S = \mathbf{x}\boldsymbol{\psi} + w$ and

$\eta_\ell = \mathbf{x}\phi + w$, respectively. The survival function for an individual is given by

$$S(t|\mathbf{x}, w) = \left[1 + \frac{\nu}{\xi} \mathcal{R}_0(t) \right]^{-\xi},$$

where $\nu = \exp(\eta_S)$ and $\xi = \exp(\eta_\ell)$.

We initiate by sampling a variable U from $U(0, 1)$. Then, define u as the survival function at time t given the covariates vector \mathbf{x} and the frailty w , expressed as $u := S(t|\mathbf{x}, w)$. Generate the time-to-event by applying the inverse of the survival function, represented by $t = S^{-1}(u|\mathbf{x}, w)$. To do this, note that

$$u = \left[1 + \frac{\nu}{\xi} \mathcal{R}_0(t) \right]^{-\xi} \Leftrightarrow \mathcal{R}_0(t) = \frac{\xi}{\nu} \left(u^{-\frac{1}{\xi}} - 1 \right).$$

But,

$$\mathcal{R}_0(t) = \exp [H_0(t)] - 1.$$

In this way, we can write

$$\begin{aligned} \frac{\xi}{\nu} \left(u^{-\frac{1}{\xi}} - 1 \right) &= \exp [H_0(t)] - 1 \\ \Leftrightarrow H_0(t) &= \log \left[\frac{\xi}{\nu} \left(u^{-\frac{1}{\xi}} - 1 \right) + 1 \right]. \end{aligned} \quad (4.1)$$

Define $\Omega := \frac{\xi}{\nu} \left(u^{-\frac{1}{\xi}} - 1 \right)$. So, we can rewrite (4.1) as

$$\vartheta := H_0(t) = \log(\Omega + 1). \quad (4.2)$$

The function $H_0(t)$ depends on the baseline chosen for data generation. Therefore, to calculate the time-to-event, we can alternatively utilize the inverse of the baseline cumulative hazard function, i.e., $t = H_0^{-1}(\vartheta)$. We choose an exponential baseline, whose rate parameter is γ . In this case, the cumulative hazard function is given by

$$H_0(t) = \frac{t}{\gamma}. \quad (4.3)$$

Thus, we can obtain the time-to-event by inverting the function presented in (4.3), whose result is

$$t = \gamma\vartheta, \quad (4.4)$$

in which ϑ is obtain in (4.2).

Additionally, for each individual, we generate the administrative censoring time as a random variable C , drawn from $U(0, \max f_u)$, where $\max f_u$ is defined previously and represents the maximum follow-up time. The follow-up time of an individual is given by

$$y = \min\{t, c\}, \quad (4.5)$$

and the failure state indicator of this time is denoted by $\delta = I\{t \leq c\}$, where $\delta = 1$, y represents a failure time, whereas $\delta = 0$, y corresponds to a right-censored time.

From the simulation study, aiming to compare the performance of our models, we are interested in some statistics. Consider a generic parameter, whose true value is Φ , and $\hat{\Phi}_K$ is the posterior estimate obtained from the k -th Monte Carlo replica, which $k \in \{1, \dots, M_C\}$. The average estimate (est) is given by

$$\text{est}(\Phi) = \frac{1}{M_C} \sum_{k=1}^{M_C} \hat{\Phi}_k.$$

We compute bias as $b_k(\Phi) = \hat{\Phi}_k - \Phi$ and the relative biases (RB) as

$$\text{RB}(\%) = 100 \times \frac{1}{M_C} \sum_{k=1}^{M_C} \frac{b_k}{|\Phi|}.$$

Additionally, we can compute the average standard error (ASE) of the estimates by

$$\text{ASE} = \frac{1}{M_C} \sum_{k=1}^{M_C} \text{se}(\hat{\Phi}_k),$$

where $\text{se}(\hat{\Phi}_k)$ represents the standard error estimates of Φ . We are also interested in evaluating the standard deviation estimate (SDE) of Φ by

$$\text{SDE} = \sqrt{\frac{1}{M_C - 1} \sum_{k=1}^{M_C} [\hat{\Phi}_k - \text{est}(\Phi)]^2}.$$

In a well-fitted model, we note these characteristics: $\text{est}(\Phi)$ should be close to the true value Φ ; SDE and ASE should be similar; $\text{RB}(\%)$ should approximate zero; and CP should be close to the pre-defined confidence level $(1 - \alpha)$. When the $\text{ASE} < \text{SDE}$, is expected $\text{CP} < 1 - \alpha$. On the other hand, if $\text{ASE} > \text{SDE}$, it is expected that $\text{CP} > 1 - \alpha$.

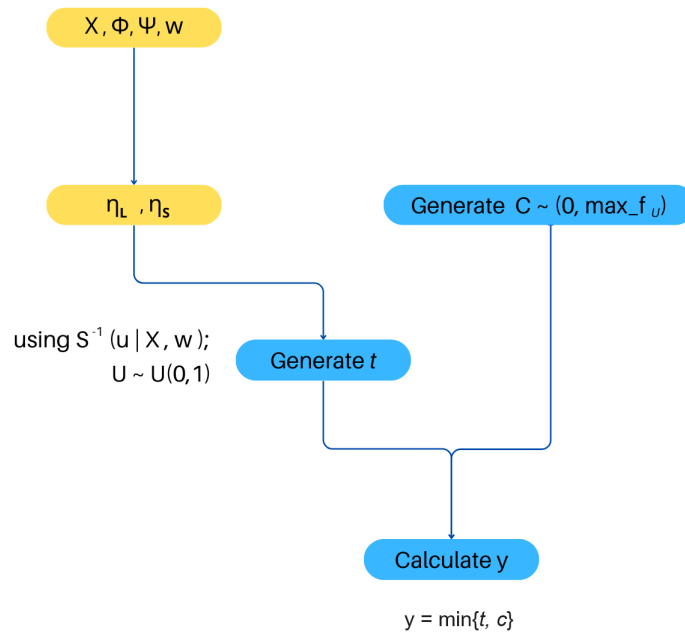
In this simulation study, we also evaluated the lower (LW) and upper (UP) limits of the 95% credible intervals. These limits are calculated from the 2.5% and 97.5% quartiles of the parameter's posterior density functions.

Before starting to detail the simulation study scenarios, it is necessary to highlight that in all of them, the data are generated using the YP model with an exponential baseline. Furthermore, we consider that $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim \text{Normal}(0, 1)$. The individual frailty w was generated from $\text{Normal}(0, 1)$. The values are the same as those chosen by [Li et al. \(2019\)](#), whose work has influenced some aspects of this thesis. We generated $M_C = 250$ replicas, each one with $L = 300$ individuals. Furthermore, regarding parameter estimation, for the piecewise exponential baseline, we chose the number of intervals $m = 5$. In all simulations, we set $\alpha = 0.05$.

4.1 Analysis of the Yang and Prentice frailty models (Class 1)

This section aims to detail the configurations employed in our simulation study in which we evaluate the first class of YP models. In the dataset generated in Section 4.1.1, each individual can experience an event only once. In this way, individual frailties are utilized to capture unobserved heterogeneities. Therefore, the time-to-event t_i of the individual i is generated conditionally to an individual frailty element w_i . Figure 4.1 shows a schematic representation of the generation of times to events considering individual frailties (class 1).

Figure 4.1: Schematic representation of the generation of the times to event considering individual frailties (class 1)

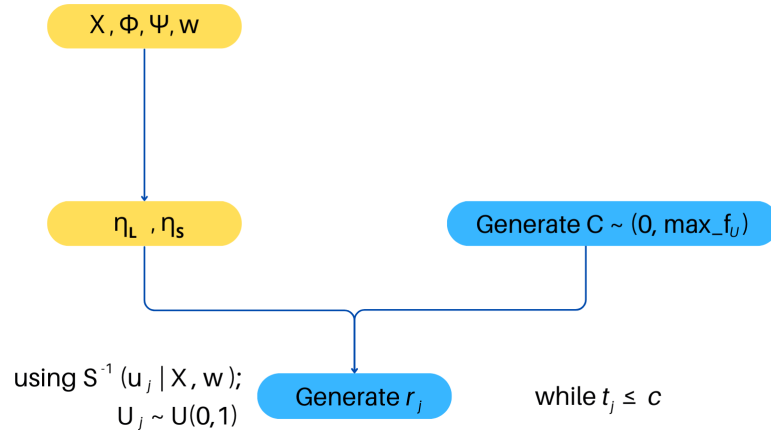


Source: Prepared by the author.

Conversely, in Section 4.1.2, the datasets simulated represent individuals who may experiment recurrent events. Thus, each individual can be considered a cluster, in which shared frailties are employed also to model the correlation among recurrence times. We generate the gap-times $r_{i,j}$ between the recurrences $j-1$ and j of the i -th individual. This generation is done using $r_{i,j} = \gamma\vartheta$, with $\vartheta = H_0(r_{i,j})$, in which $H_0(\cdot)$ is defined in (4.2). The gap-time $r_{i,j}$ is generated conditionally to frailty w_i of the i -th individual. Thus, only one value w_i is generated for each i . The values $r_{i,j}$ are generated interactively until $t_{i,j} = \sum_{j'=1}^j r_{i,j'} \leq c_i$. Figure 4.2 presents a schematic representation of the generation of the gap times between recurrent events considering shared frailties (class 1) for an

individual.

Figure 4.2: Schematic representation of the generation of the gap times between recurrent events considering shared frailties (class 1)



Source: Prepared by the author.

The true values of the parameters established for simulations related to the first class of models are in Table 4.1.

Table 4.1: True values

ϕ_1	ϕ_2	ψ_1	ψ_2	γ	σ_w
-1.0	2.0	2.0	2.0	1.2	1.0

The following prior distributions were used:

$$\psi \sim \text{Normal}(0, 4^2),$$

$$\phi \sim \text{Normal}(0, 4^2),$$

and

$$\gamma \sim \text{LogNormal}(0, 2).$$

We assume that the standard deviation of the frailty is

$$\sigma_w \sim \text{Gamma}(0.1, 0.1).$$

The prior distributions established for the regression coefficients ψ , ϕ , and the standard deviation of the frailty σ_w are weakly informative (Stan Development Team, 2023). A weakly informative prior is designed such that, in the presence of a sufficiently large dataset, the likelihood will dominate the estimation rendering the prior's influence relatively insignificant (Gabry et al., 2019). For the parameter γ , we choose the prior distribution that provides greater stability in the inferential process, as suggested by Demarqui et al. (2019). All simulation study results, shown in this section, are also available online¹.

¹Access the link cassiushenrique.shinyapps.io/appSimulationsFrailty.

Table 4.2: Monte Carlo summary statistics of the YP_{EX} , YP_{PE} , and YP_{BP} models with individual frailties, for $L = 300$ and $M_C = 250$.

fitted model	par	true	est	RB (%)	ASE	SDE	95% CI		
							LW	UP	CP
YP_{EX}	ϕ_1	-1	-0.9406	5.9395	0.2970	0.2935	-1.4732	-0.3098	0.9360
	ϕ_2	2	2.0795	3.9738	0.3404	0.3388	1.4428	2.7755	0.9440
	ψ_1	2	2.0175	0.8726	0.2941	0.2940	1.4410	2.5919	0.9560
	ψ_2	2	2.0094	0.4704	0.1713	0.1713	1.6731	2.3453	0.9640
	σ_w	1	0.9971	-0.2909	0.1763	0.1763	0.6489	1.3418	0.9400
YP_{PE}	ϕ_1	-1	-0.9399	6.0081	0.3085	0.3049	-1.4925	-0.2813	0.9520
	ϕ_2	2	2.0681	3.4029	0.4606	0.4595	1.2328	3.0187	0.9360
	ψ_1	2	2.1097	5.4830	0.3482	0.3452	1.4506	2.8150	0.9600
	ψ_2	2	2.0802	4.0083	0.2497	0.2481	1.6254	2.6026	0.9360
	σ_w	1	0.9944	-0.5598	0.3186	0.3185	0.4260	1.6263	0.9320
YP_{BP}	ϕ_1	-1	-0.9632	3.6782	0.3076	0.3063	-1.5138	-0.3090	0.9520
	ϕ_2	2	2.1899	9.4949	0.5176	0.5086	1.2725	3.2830	0.9240
	ψ_1	2	2.0925	4.6234	0.3439	0.3418	1.4595	2.8115	0.9640
	ψ_2	2	2.0851	4.2546	0.2574	0.2556	1.6606	2.6741	0.9640
	σ_w	1	1.0628	6.2832	0.3346	0.3306	0.4848	1.7816	0.9440

4.1.1 Yang and Prentice model with individual frailty

In this section, we present the results of the Monte Carlo study for YP_{EX} , YP_{PE} , and YP_{BP} models with individual frailties ($n_i = 1, \forall i \in \{1, \dots, L\}$). In simulated data, approximately 68.4% of the individuals experienced the event of interest, on average. The Monte Carlo summary statistics in Table 4.2 compare the performance of our models. Regarding the estimation of the parameters for the YP_{BP} model, we consider $m = L^{0.4}$, where L is the total number of individuals. The choice of the polynomial degree is motivated by the suggestion of [Osman and Ghosh \(2012\)](#).

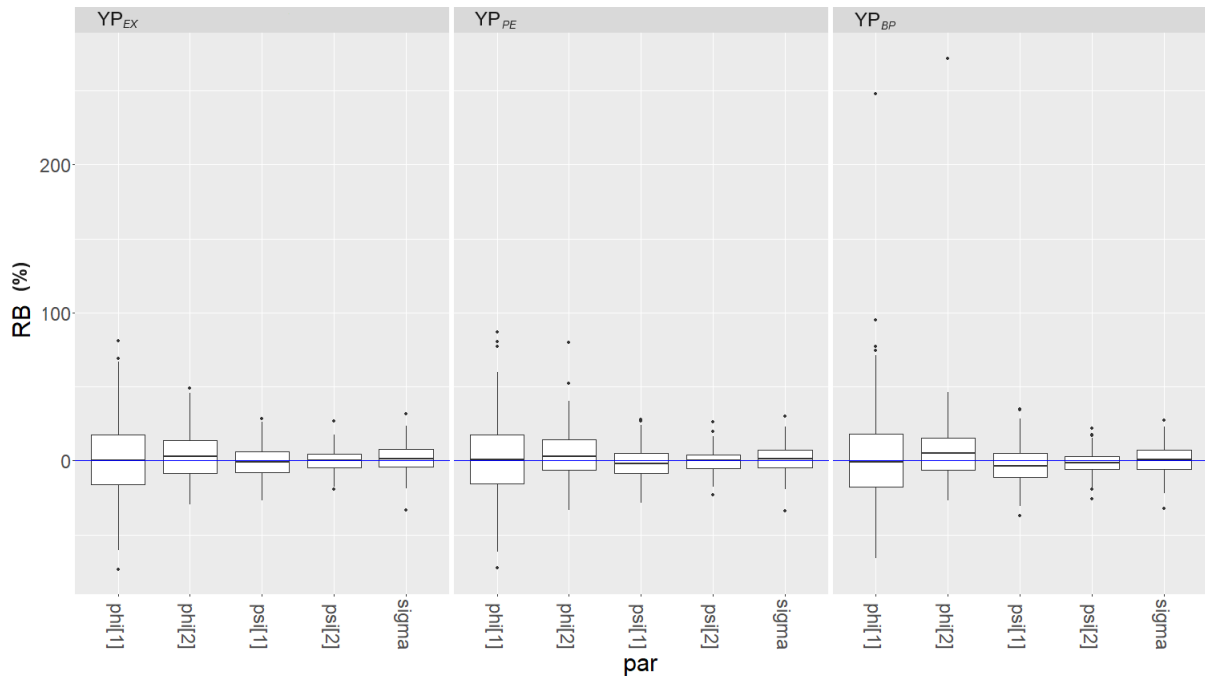
The estimates from our models are close to the true values. The estimates of σ_w and ϕ_2 by the YP_{BP} model are slightly less accurate, deviating more from their true values, compared to the corresponding estimates obtained from the other models. The YP_{EX} model presented smaller biases for the short-term effect coefficients (ψ_1 and ψ_2). The YP_{EX} and YP_{PE} models show greater RB for the parameters of dichotomous variables in comparison to continuous covariates effects. For all the models, credible intervals are similar. Furthermore, the ASE values are close to SDE, and CP values are close to the desired level 0.95, indicating good performance. We also evaluated the MCMC chains generated by our models and observed a satisfactory convergence of all parameters in the simulated study.

Overall, all models show good performance with some variations in precision and

bias with YP_{BP} and YP_{PE} providing similar results to the parametric model that generated data. Thus, the choice of the model would depend on the importance of each parameter, and the balance between bias and precision as indicated by these summary statistics.

Figure 4.3 presents the boxplot of the relative biases of the estimates of each parameter of the three models. The medians closest to zero were found by ϕ_1 in the YP_{BP} model, ψ_1 , ψ_2 and σ_w in the YP_{EX} model and by ψ_2 in the YP_{PE} model. The variability of the estimates of each parameter is not so different when we look at the three models, although the YP_{EX} presents less variability in the short-term regression coefficients and the frailty standard deviation. This is well-accepted since it equals the generating model. The regression coefficients of the long-term effect showed a greater number of outliers. However, the YP_{BP} model also has more outliers in the other parameters. We can state that all models performed well as they have median biases close to zero for all parameters.

Figure 4.3: Boxplot of RB(%) for the YP_{EX} , YP_{PE} , and YP_{BP} models with individual frailties, for $L = 300$ and $M_C = 250$.



Source: Prepared by the author.

4.1.2 Yang and Prentice model with shared frailty

We continue to evaluate the models that comprise our first class. We also carried out a Monte Carlo study considering that each individual can have various recurrences

Table 4.3: Monte Carlo summary statistics of the YP_{EX} , YP_{PE} , and YP_{BP} models with shared frailties, for $L = 300$ and $M_C = 250$.

fitted model	par	true	est	RB (%)	ASE	SDE	95% CI		
							LW	UP	CP
YP_{EX}	ϕ_1	-1	-0.9831	1.6948	0.2586	0.2583	-1.4664	-0.4492	0.9760
	ϕ_2	2	2.0664	3.3213	0.2951	0.2940	1.5314	2.6889	0.9560
	ψ_1	2	1.9834	-0.8318	0.2090	0.2090	1.5739	2.3930	0.9640
	ψ_2	2	2.0076	0.3822	0.1377	0.1377	1.7418	2.2819	0.9520
	σ_w	1	1.0172	1.7159	0.0911	0.0908	0.8519	1.2085	0.9600
YP_{PE}	ϕ_1	-1	-0.9692	3.0824	0.2676	0.2666	-1.4629	-0.4144	0.9760
	ϕ_2	2	2.0885	4.4248	0.3237	0.3218	1.5101	2.7747	0.9640
	ψ_1	2	1.9628	-1.8625	0.2128	0.2125	1.5461	2.3796	0.9600
	ψ_2	2	1.9958	-0.2118	0.1387	0.1386	1.7282	2.2718	0.9440
	σ_w	1	1.0133	1.3316	0.0906	0.0904	0.8491	1.2037	0.9600
YP_{BP}	ϕ_1	-1	-0.9755	2.4497	0.3115	0.3109	-1.5326	-0.3207	0.9560
	ϕ_2	2	2.1264	6.3217	0.3226	0.3186	1.5465	2.8140	0.9640
	ψ_1	2	1.9396	-3.0183	0.2435	0.2426	1.4703	2.4252	0.9360
	ψ_2	2	1.9755	-1.2260	0.1447	0.1446	1.6976	2.2649	0.9440
	σ_w	1	1.0099	0.9855	0.0938	0.0937	0.8401	1.2073	0.9720

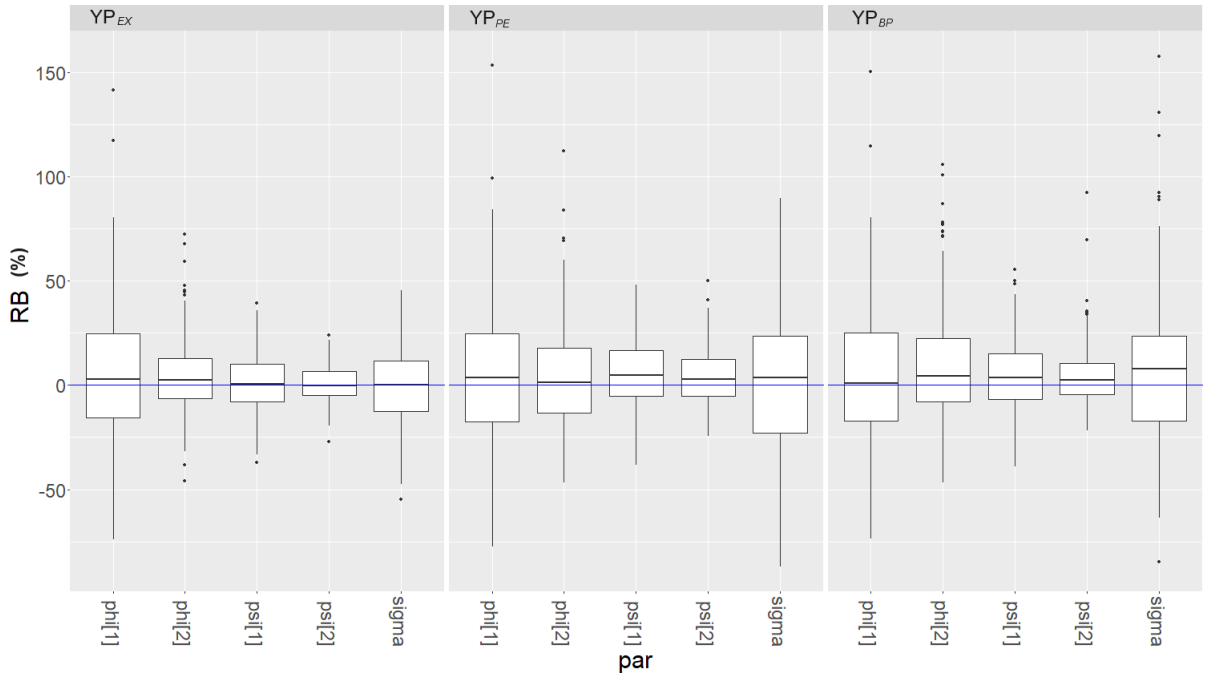
($n_i \geq 1, \forall i \in \{1, \dots, L\}$). We fit the YP_{EX} , YP_{PE} , and YP_{BP} models with shared frailty (because each individual is a cluster). As suggested by [Osman and Ghosh \(2012\)](#), in the YP_{BP} , we consider $m = \left[\sum_{i=1}^L n_i \right]^{0.4}$, in which $\sum_{i=1}^L n_i$ is the total number of recurrent events. From [Table 4.3](#), some interesting comparisons emerge.

As expected, the YP_{EX} shows a high degree of accuracy in estimating the parameters. The estimates for all the parameters are very close to their true values, with relative biases (RB) ranging from -0.8318% to 3.3213% . In contrast, when the YP_{PE} is fitted, the relative biases for the same parameters increase slightly, in magnitude, except for σ_w . For instance, the estimate for ϕ_1 shows a relative bias of 3.0824% , compared to 1.6948% in the YP_{EX} model. However, these biases remain within acceptable limits. The YP_{BP} demonstrates a further increase in relative biases for parameters such as ϕ_2 , where the bias reaches 6.3217% , the highest among the three models. Additionally, The ASE and SDE estimates are close to each other. The CP values are approximately 0.95, deviating by no more than 0.026 from this level in all cases.

Notably, across all models, the estimation of the frailty parameter σ_w is accurate, with relative biases under 2%. This indicates that all three models are reliable in capturing the shared frailty component, which is an essential aspect.

Overall, our models show similar performance in terms of the criteria analyzed. Some of the comparisons discussed can also be evaluated from [Figure 4.4](#).

Figure 4.4: Boxplot of RB(%) for the YP_{EX} , YP_{PE} , and YP_{BP} models with shared frailties, for $L = 300$ and $M_C = 250$.



Source: Prepared by the author.

4.2 Analysis of the joint frailty-copula models (Class 2)

In this section, we discuss the outcomes of a simulation study conducted to assess the performance of our joint frailty-copula models. Consider the scenario in which we evaluate the time to the terminal event for n individuals. Each one can experience recurrent events. Some characteristics that supposedly influence the time to terminal events and recurrent events were recorded at the beginning of their follow-up period. We denote, for an individual, these characteristics as $1 \times p$ vectors $\mathbf{x}^{(T)}$ and $\mathbf{x}^{(R)}$, respectively.

We define the short-term ($\boldsymbol{\psi}^{(T)}$ and $\boldsymbol{\psi}^{(R)}$) and long-term ($\boldsymbol{\phi}^{(T)}$ and $\boldsymbol{\phi}^{(R)}$) regression coefficients as $1 \times p$ vectors. Furthermore, we specify the variance σ_w^2 of the individual frailty. In our models, we assume $w \sim \text{Normal}(0, \sigma_w^2)$. The baseline function parameters $\gamma^{(T)}$ and $\gamma^{(R)}$ are also established. Lastly, we choose the value of the association parameter of copula θ .

We calculate the linear predictors of our models. The short-term linear predictors of terminal events and recurrent events for an individual are given by, respectively $\eta_S^{(T)} = \mathbf{x}^{(T)}\boldsymbol{\psi}^{(T)} + w$, and $\eta_S^{(R)} = \mathbf{x}^{(R)}\boldsymbol{\psi}^{(R)} + w$. Moreover, the long-term linear predictors of terminal events and recurrent events for an individual, respectively $\eta_\ell^{(T)} = \mathbf{x}^{(T)}\boldsymbol{\phi}^{(T)} + w$, and $\eta_\ell^{(R)} = \mathbf{x}^{(R)}\boldsymbol{\phi}^{(R)} + w$,

We can obtain the PH and PO models from the Yang and Prentice (YP) model. To obtain the PH model, do $\eta_\ell^{(T)} = \eta_S^{(T)}$, and $\eta_\ell^{(R)} = \eta_S^{(R)}$. In contrast, to obtain the PO model, the long-term linear predictors $\eta_\ell^{(T)}$ and $\eta_\ell^{(R)}$ are defined with $\eta_\ell^{(T)} = 0$, and $\eta_\ell^{(R)} = 0$. Using these quantities, we generate the time to the terminal event by the inverse $t = S_D^{-1}(u|\mathbf{x}, w)$ which is equivalent to the result shown in (4.4) and the observed follow-up time as shown in (4.5).

We now proceed to the generation of recurrent events given the times to the terminal event. We set j as the recurrent event count for the i -th individual, with $j \in \{0, 1, \dots, n_i\}$. The time between recurrent events is generated by the inverse $r_{i,j} = S_R^{-1}(v_{i,j}|\mathbf{x}, w)$ and it can be calculated using (4.4). We define $U_{i,j} := C(u_i, v_{i,j}; \theta)$, where $u_i = S_D(t_i|\mathbf{x}_i, w_i)$ and $v_{i,j} = S_R(r_{i,j}|\mathbf{x}_i, w_i)$. The θ parameter takes a unique value for all individuals. Generate $U_{i,j} \sim U(0, 1)$. Then, we can find the values of the survival function $v_{i,j}|u_{i,j}$ of the j -th recurrent event of an individual through the inverse of the copula function $C_{(10)}^{-1}(u_i, U_{i,j}; \theta)$, according to Nelsen (2006).

Assume that the joint distribution of u_i and $v_{i,j}$ is a Clayton copula whose density is

$$C(u_i, v_{i,j}, \theta) = (u_i^{-\theta} + v_{i,j}^{-\theta} - 1)^{-\frac{1}{\theta}}.$$

The derivative of this copula with respect to u is given by

$$C_{(10)}(u_i, v_{i,j}, \theta) = u_i^{-(\theta+1)} (u_i^{-\theta} + v_{i,j}^{-\theta} - 1)^{-\frac{\theta+1}{\theta}}. \quad (4.6)$$

Thus, the inverse values of the Clayton copula function are determined as follows:

$$v_{i,j} = C_{(10)}^{-1}(u_i, U_{i,j}, \theta) = \left[\left(U_{i,j}^{-\frac{\theta}{\theta+1}} - 1 \right) u_i^{-\theta} + 1 \right]^{-\frac{1}{\theta}}. \quad (4.7)$$

Using

$$\Omega_{i,j} = \frac{\xi_i^{(R)}}{\nu_i^{(R)}} \left(\frac{-\frac{1}{\xi_i^{(R)}}}{v_{i,j}} - 1 \right),$$

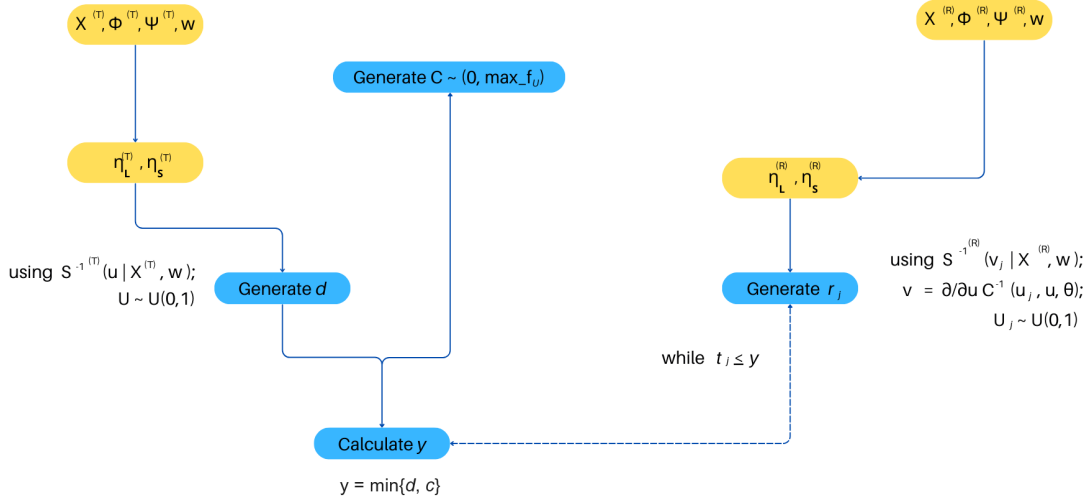
we define $\vartheta_{i,j} := H_0(r_{i,j}) = \log(\Omega_{i,j} + 1)$. Thus, if the baseline is based on the exponential distribution, then

$$r_{i,j} = \gamma^{(R)} \vartheta_{i,j}.$$

We compute the times between recurrent events $r_{i,j}$ while $t_{i,j} = \sum_{j'=1}^j r_{i,j'} \leq y_i$. If $j = n_i + 1$, $r_{i,n_i+1} = r_i - \sum_{j=1}^{n_i} r_{i,j}$ is the gap-time between the n_i -th recurrent event and the end of follow-up.

It is worth highlighting that the frailty used to generate the times is computed at the individual level. Therefore, for the times until the terminal event and the gap times between recurrent events of the same individual, a unique value of w is generated. Figure 4.5 shows a schematic representation of the generation of times until the terminal event and the gap times between recurrent events using the Clayton copula (class 2).

Figure 4.5: Schematic representation of the generation of times until the terminal event and the gap times between recurrent events using the Clayton copula (class 2)



Source: Prepared by the author.

Now, we present the results from the Monte Carlo study of the joint frailty-copula models. The true values of the parameters of our models are in Table 4.4. To simplify the notation, from now on, we assume $\beta^{(T)} = \phi^{(T)} = \psi^{(T)}$ and $\beta^{(R)} = \phi^{(R)} = \psi^{(R)}$, in PH models. Moreover, in PO models, we consider $\beta^{(T)} = \psi^{(T)}$ and $\beta^{(R)} = \psi^{(R)}$, since, under this model, $\phi^{(T)} = \psi^{(T)} = \mathbf{0}$. In YP models, we still use the notation established in this section.

Table 4.4: True values

	PH	PO	YP
$\phi_1^{(R)}$	2.00	0.00	-1.00
$\phi_2^{(R)}$	2.00	0.00	2.00
$\phi_1^{(T)}$	1.00	0.00	-1.00
$\phi_2^{(T)}$	1.00	0.00	1.00
$\psi_1^{(R)}$	2.00	2.00	2.00
$\psi_2^{(R)}$	2.00	2.00	2.00
$\psi_1^{(T)}$	1.00	1.00	1.00
$\psi_2^{(T)}$	1.00	1.00	1.00
$\gamma^{(R)}$	1.20	1.20	1.20
$\gamma^{(T)}$	1.00	1.00	1.00
σ_w	1.00	1.00	1.00
τ_κ	0.25	0.25	0.25

We selected a τ_κ value of 0.25 based on the observation that the real data from our application, discussed in Section 5.2, exhibit a notably weak correlation between recurrent and terminal events. In generated datasets, the individuals had, on average, 2.48 recurrent

events, with a standard deviation of 5.74. Approximately 76.5% of individuals experienced a terminal event.

The following prior distributions were used:

$$\begin{aligned}\boldsymbol{\beta}^{(T)} &\sim \text{Normal}(0, 4^2), \\ \boldsymbol{\beta}^{(R)} &\sim \text{Normal}(0, 4^2), \\ \boldsymbol{\psi}^{(T)} &\sim \text{Normal}(0, 4^2), \\ \boldsymbol{\psi}^{(R)} &\sim \text{Normal}(0, 4^2), \\ \boldsymbol{\phi}^{(T)} &\sim \text{Normal}(0, 4^2), \\ \boldsymbol{\phi}^{(R)} &\sim \text{Normal}(0, 4^2), \\ \theta &\sim \text{Gamma}(0.01, 0.01), \\ \gamma^{(T)} &\sim \text{LogNormal}(0, 2),\end{aligned}$$

and

$$\gamma^{(R)} \sim \text{LogNormal}(0, 2).$$

We assume for the standard deviation of the frailty the prior

$$\sigma_w \sim \text{Gamma}(0.1, 0.1).$$

The prior distributions established for the regression coefficients $\boldsymbol{\beta}^{(T)}$, $\boldsymbol{\beta}^{(R)}$, $\boldsymbol{\psi}^{(T)}$, $\boldsymbol{\psi}^{(R)}$, $\boldsymbol{\phi}^{(T)}$, and $\boldsymbol{\phi}^{(R)}$, σ_w , and the copula association parameter θ are weakly informative. For the parameters $\gamma^{(R)}$ and $\gamma^{(T)}$, we choose prior distributions that provide greater stability in the inferential process, as suggested by [Demarqui et al. \(2019\)](#).

For the piecewise exponential and Bernstein baselines we established m in the same way as the Sections 4.1.1 and 4.2 for the terminal and recurring events, respectively. The results found in the Monte Carlo study are presented in Tables 4.5, 4.6, and 4.7. Analysis of Table 4.5 revealed interesting results. While the PH_{EX} model exhibited the smallest biases, the PH_{PE} showed higher relative biases for both the regression coefficients and the frailty standard deviation parameters compared to the other models. On the other hand, the PH_{BP} demonstrated competitive performance in terms of relative biases. This indicates that the model based on the Bernstein polynomials to handle the baseline functions effectively captured the complexities of the data, resulting in accurate estimates.

In terms of the width of the credible intervals, there was also remarkable similarity among the models. However, it was observed that the generator model tends to produce narrower credible intervals compared to the other models in most cases. This suggests that the generator model was able to provide more precise estimates, leading to tighter intervals. This result was expected.

Overall, these findings provide valuable insights into the performance of the different models. While the PH_{EX} exhibited the smallest biases and narrower credible intervals,

Table 4.5: Monte Carlo summary statistics of the joint frailty-copula models: PH_{EX} , PH_{PE} , and PH_{BP} when the generator is equivalent to the PH_{EX} ($L = 300$ and $M_C = 250$).

fitted model	par	true	est	RB (%)	ASE	SDE	95% CI		
							LW	UP	CP
PH_{EX}	$\beta_1^{(R)}$	2.00	1.9943	-0.2852	0.1725	0.1725	1.6638	2.3228	0.9640
	$\beta_2^{(R)}$	2.00	2.0093	0.4650	0.0969	0.0968	1.8319	2.1943	0.9360
	$\beta_1^{(T)}$	1.00	0.9800	-1.9998	0.1944	0.1940	0.6151	1.3439	0.9560
	$\beta_2^{(T)}$	1.00	1.0035	0.3515	0.1000	0.1000	0.8174	1.1930	0.9440
	σ_w	1.00	1.0360	3.6011	0.1241	0.1228	0.8631	1.2729	0.9640
	τ_κ	0.25	0.2447	-2.1384	0.0322	0.0317	0.1852	0.2995	0.9480
PH_{PE}	$\beta_1^{(R)}$	2.00	2.0684	3.4204	0.1860	0.1848	1.7174	2.4380	0.9520
	$\beta_2^{(R)}$	2.00	2.0395	1.9731	0.1213	0.1209	1.8173	2.2795	0.9400
	$\beta_1^{(T)}$	1.00	1.0516	5.1577	0.1873	0.1846	0.6948	1.4205	0.9280
	$\beta_2^{(T)}$	1.00	1.0258	2.5754	0.1075	0.1068	0.8260	1.2387	0.9480
	σ_w	1.00	1.0378	3.7819	0.1159	0.1145	0.8399	1.2681	0.9400
	τ_κ	0.25	0.2434	-2.6301	0.0334	0.0327	0.1795	0.3075	0.9240
PH_{BP}	$\beta_1^{(R)}$	2.00	1.9934	-0.3297	0.1820	0.1819	1.6516	2.3525	0.9560
	$\beta_2^{(R)}$	2.00	2.0261	1.3065	0.1140	0.1138	1.8247	2.2535	0.9600
	$\beta_1^{(T)}$	1.00	0.9774	-2.2645	0.1896	0.1891	0.6176	1.3448	0.9520
	$\beta_2^{(T)}$	1.00	1.0168	1.6796	0.1031	0.1028	0.8231	1.2218	0.9360
	σ_w	1.00	1.0369	3.6852	0.1156	0.1143	0.8509	1.2694	0.9600
	τ_κ	0.25	0.2418	-3.2749	0.0301	0.0290	0.1840	0.2977	0.9240

the PH_{BP} showed competitive performance and captured the complexities of the data well. The PH_{PE} , on the other hand, exhibited higher biases and performed relatively less favorably. These results highlight the importance of carefully selecting the appropriate model structure to obtain accurate and reliable parameter estimates in multivariate survival analysis.

Upon further analysis of PO family models by Table 4.6, we made some observations. The generator PO_{EX} exhibited the smallest biases. On the other hand, the PO_{PE} showed higher relative biases for all coefficients. The PO_{BP} demonstrated competitive performance in terms of RB. It performed similarly to the true model and showed superior performance compared to the PO_{PE} . This suggests that the use of Bernstein polynomials for modeling the baseline functions provided certain advantages, capturing the complexities of the data more effectively.

Regarding the amplitudes of the credible intervals, it was observed that the generator model tends to produce more restricted intervals compared to the other models. However, the use of Bernstein polynomials in the PO_{BP} conferred certain advantages in terms of providing precise and narrower estimates, in comparison with PO_{PE} . The tighter intervals generated by the PO_{BP} imply a higher precision in the estimated parameter val-

Table 4.6: Monte Carlo summary statistics of the joint frailty-copula models: PO_{EX} , PO_{PE} , and PO_{BP} when the generator is equivalent to the PO_{EX} ($L = 300$ and $M_C = 250$).

fitted model	par	true	est	RB (%)	ASE	SDE	95% CI		
							LW	UP	CP
PO_{EX}	$\beta_1^{(R)}$	2.00	1.9987	-0.0637	0.1930	0.1930	1.6329	2.3616	0.9520
	$\beta_2^{(R)}$	2.00	2.0021	0.1043	0.1068	0.1068	1.7997	2.2076	0.9520
	$\beta_1^{(T)}$	1.00	0.9781	-2.1920	0.2562	0.2558	0.4948	1.4141	0.9760
	$\beta_2^{(T)}$	1.00	0.9937	-0.6263	0.1346	0.1345	0.7391	1.2403	0.9240
	σ_w	1.00	1.0021	0.2148	0.1390	0.1390	0.7449	1.2674	0.9560
	τ_κ	0.25	0.2383	-4.6822	0.0328	0.0306	0.1750	0.2911	0.9520
PO_{PE}	$\beta_1^{(R)}$	2.00	2.0600	2.9980	0.2162	0.2153	1.6518	2.4832	0.9400
	$\beta_2^{(R)}$	2.00	2.0235	1.1725	0.1326	0.1324	1.7731	2.2863	0.9360
	$\beta_1^{(T)}$	1.00	1.0765	7.6518	0.2441	0.2383	0.6099	1.5525	0.9360
	$\beta_2^{(T)}$	1.00	1.0161	1.6121	0.1345	0.1342	0.7606	1.2799	0.9000
	σ_w	1.00	1.0170	1.6994	0.1814	0.1811	0.6663	1.3659	0.9560
	τ_κ	0.25	0.2375	-4.9986	0.0356	0.0331	0.1675	0.3022	0.9600
PO_{BP}	$\beta_1^{(R)}$	2.00	2.0096	0.4796	0.2310	0.2309	1.5848	2.4628	0.9360
	$\beta_2^{(R)}$	2.00	2.0131	0.6567	0.1243	0.1243	1.7864	2.2611	0.9520
	$\beta_1^{(T)}$	1.00	1.0227	2.2710	0.2490	0.2485	0.5488	1.5103	0.9440
	$\beta_2^{(T)}$	1.00	1.0150	1.4955	0.1319	0.1317	0.7679	1.2736	0.9240
	σ_w	1.00	1.0053	0.5324	0.1623	0.1623	0.7062	1.3249	0.9320
	τ_κ	0.25	0.2407	-3.7274	0.0306	0.0292	0.1785	0.2954	0.9520

ues, which can be beneficial in various inferential analyses.

Regarding the YP family of models, some observations can be made. Their results are shown in Table 4.7. First, the estimates of the YP_{EX} exhibited less bias in most parameters compared to the other models. However, it is worth noting that the coefficients of Bernoulli's random variable showed relatively higher biases. It is important to consider that outliers in the posterior samples can influence this quantity. To further assess these biases, we will examine their medians in subsequent analyses. When considering the remaining parameters, the YP_{BP} demonstrated competitive performance with similar relative biases compared to the other models.

For all scenarios discussed in this section, the values of ASE and SDE are similar. In terms of the coverage probabilities of the credible intervals, all the models also demonstrated relatively similar performance, closely aligning with the nominal level of 95%, deviating by a maximum of 5%.

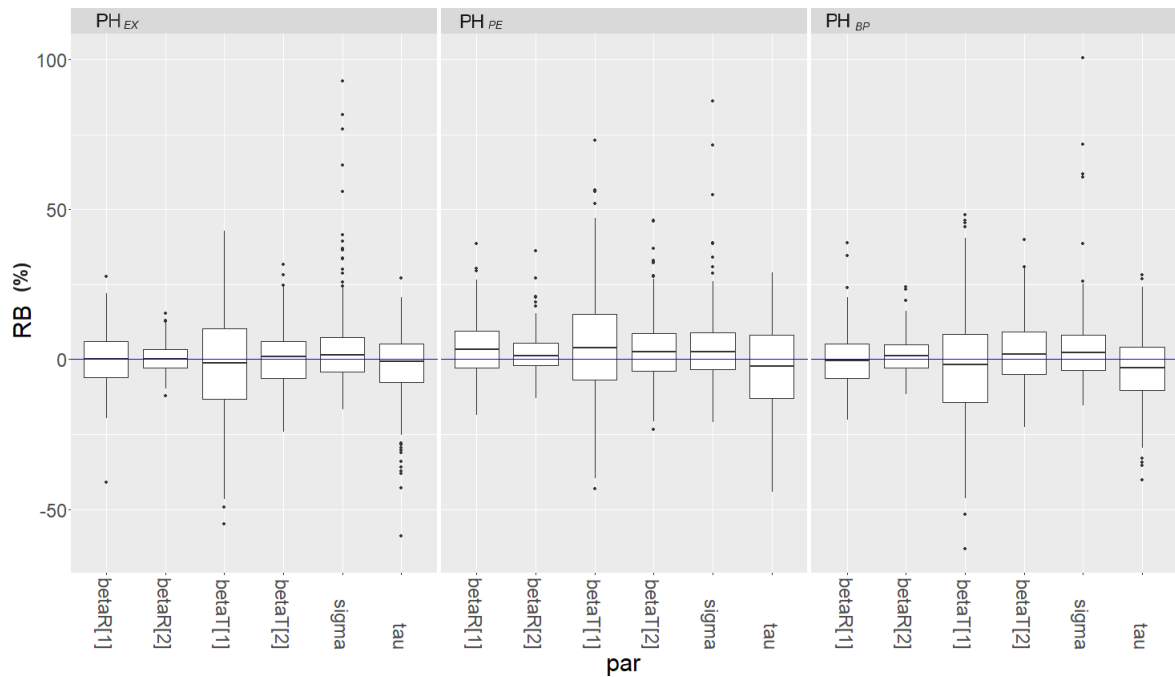
Let us now examine the boxplot depicting the relative biases of the parameters for each regression family. Figure 4.6 provides a visual representation of the performance of the PH family. Upon observing the boxplots, it is evident that all baseline functions yielded relative biases with medians close to zero. However, the estimates of σ_w displayed

Table 4.7: Monte Carlo summary statistics of the joint frailty-copula models: YP_{EX} , YP_{PE} , and YP_{BP} when the generator is equivalent to the YP_{EX} ($L = 300$ and $M_C = 250$).

fitted model	par	true	est	RB (%)	ASE	SDE	95% CI		
							LW	UP	CP
YP_{EX}	$\phi_1^{(R)}$	-1.00	-0.8366	16.3379	0.4066	0.3799	-1.3378	-0.0791	0.9440
	$\phi_2^{(R)}$	2.00	2.1282	6.4114	0.2805	0.2764	1.6488	2.6839	0.9400
	$\phi_1^{(T)}$	-1.00	-0.8874	11.2628	0.3583	0.3457	-1.3698	-0.2105	0.9560
	$\phi_2^{(T)}$	1.00	1.0366	3.6566	0.1974	0.1961	0.7031	1.4073	0.9360
	$\psi_1^{(R)}$	2.00	1.9858	-0.7123	0.2019	0.2018	1.5945	2.3712	0.9520
	$\psi_2^{(R)}$	2.00	2.0048	0.2401	0.1238	0.1238	1.7753	2.2452	0.9760
	$\psi_1^{(T)}$	1.00	0.9682	-3.1781	0.2582	0.2572	0.4702	1.4512	0.9320
	$\psi_2^{(T)}$	1.00	1.0040	0.3966	0.1407	0.1407	0.7344	1.2754	0.9640
	σ_w	1.00	1.0276	2.7611	0.1287	0.1279	0.8269	1.2779	0.9360
	τ_κ	0.25	0.2372	-5.1068	0.0342	0.0316	0.1726	0.3013	0.9200
YP_{PE}	$\phi_1^{(R)}$	-1.00	-0.8900	10.9961	0.3105	0.2984	-1.3552	-0.2936	0.9400
	$\phi_2^{(R)}$	2.00	2.1316	6.5816	0.3064	0.3021	1.6257	2.7361	0.9400
	$\phi_1^{(T)}$	-1.00	-0.9145	8.5509	0.2921	0.2848	-1.3891	-0.3370	0.9560
	$\phi_2^{(T)}$	1.00	1.0233	2.3258	0.1925	0.1919	0.6737	1.3922	0.9560
	$\psi_1^{(R)}$	2.00	2.0400	1.9984	0.2250	0.2246	1.6074	2.4818	0.9360
	$\psi_2^{(R)}$	2.00	2.0304	1.5181	0.1365	0.1362	1.7762	2.2981	0.9680
	$\psi_1^{(T)}$	1.00	1.0286	2.8563	0.2867	0.2859	0.4830	1.5943	0.9360
	$\psi_2^{(T)}$	1.00	1.0234	2.3380	0.1514	0.1508	0.7357	1.3188	0.9680
	σ_w	1.00	1.0331	3.3075	0.1407	0.1396	0.8089	1.3082	0.9320
	τ_κ	0.25	0.2361	-5.5790	0.0379	0.0348	0.1639	0.3073	0.9400
YP_{BP}	$\phi_1^{(R)}$	-1.00	-0.9336	6.6351	0.2901	0.2857	-1.3840	-0.3836	0.9440
	$\phi_2^{(R)}$	2.00	2.1773	8.8642	0.3333	0.3254	1.6371	2.8287	0.9200
	$\phi_1^{(T)}$	-1.00	-0.9568	4.3191	0.2889	0.2870	-1.4234	-0.3969	0.9560
	$\phi_2^{(T)}$	1.00	1.0331	3.3149	0.1732	0.1721	0.7094	1.3710	0.9280
	$\psi_1^{(R)}$	2.00	1.9766	-1.1697	0.2156	0.2155	1.5666	2.3975	0.9400
	$\psi_2^{(R)}$	2.00	2.0039	0.1934	0.1359	0.1359	1.7575	2.2716	0.9560
	$\psi_1^{(T)}$	1.00	0.9663	-3.3733	0.2812	0.2801	0.4311	1.5178	0.9440
	$\psi_2^{(T)}$	1.00	1.0051	0.5114	0.1553	0.1553	0.7173	1.3086	0.9720
	σ_w	1.00	1.0293	2.9257	0.1399	0.1390	0.7999	1.3060	0.9320
	τ_κ	0.25	0.2370	-5.2038	0.0361	0.0334	0.1688	0.3056	0.9160

more asymmetric biases compared to the other parameters, with a higher number of outliers across all three models. Here, an outlier is defined as a value that deviates from the nearest quartile by at least 1.5 times the interquartile range. The Kendall correlation coefficient, τ_κ , exhibited negative medians of RB in all three models, although the values were close to zero. The variabilities of the biases for each parameter were relatively similar across the three models.

Figure 4.6: Boxplot of RB(%) for the joint frailty-copula models: PH_{EX} , PH_{PE} when the generator is equivalent to the PH_{EX} model ($L = 300$ and $M_C = 250$).

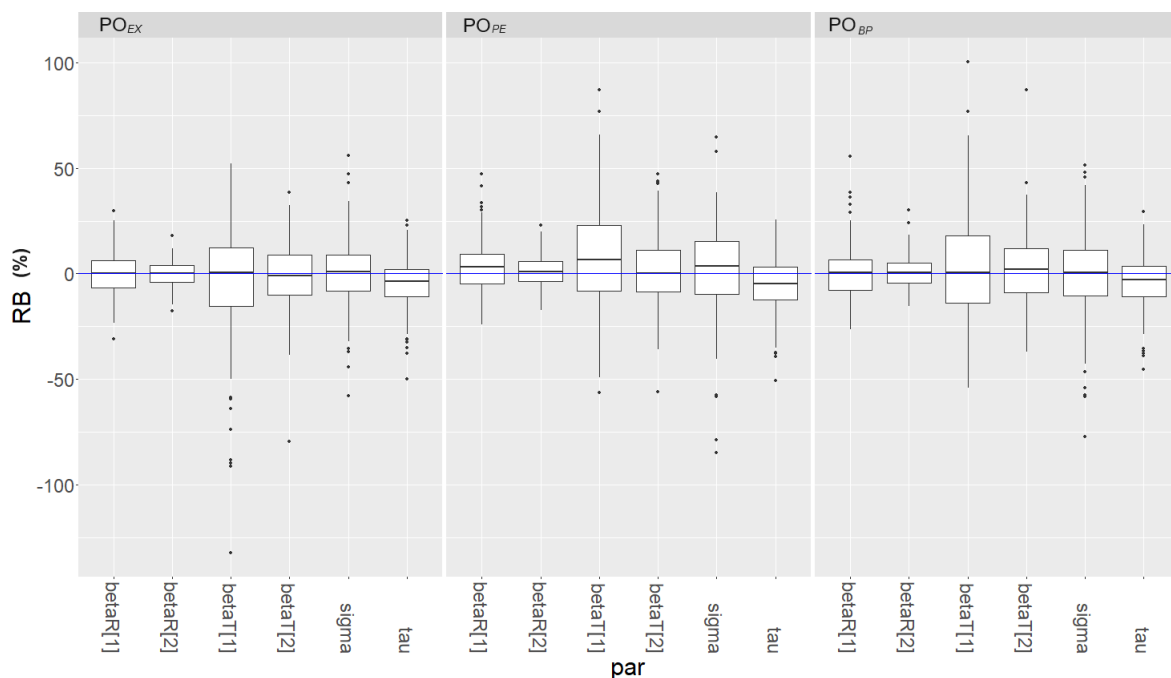


Source: Prepared by the author.

Now let's examine the performance of the relative biases for the PO regression family. Figure 4.7 provides the boxplots of the relative biases for the PO_{EX} , PO_{BP} , and PO_{PE} . Similar to the PH family, in this family all baseline functions resulted in relative biases with medians close to zero. Additionally, all parameters exhibited approximately symmetric biases. The variabilities of the biases for each parameter are similar across the three models.

Let us now delve into the analysis of the YP regression family, which exhibits a slightly different behavior compared to the PH and PO model families. Figure 4.8 displays the boxplots of the relative biases for the models with Yang and Prentice regression structure. What stands out in the YP regression family is the presence of higher mean biases for certain parameters compared to others. When examining the boxplots of the short-term regression coefficients, we observe a more significant number of outliers. This may be attributed to specific characteristics of the generated Monte Carlo replicas. These outliers have a substantial impact on the mean relative biases of certain parameters. Notably, in the YP_{EX} , even though it is the generator model, the coefficients $\phi_1^{(R)}$, $\phi_2^{(R)}$,

Figure 4.7: Boxplot of RB(%) for the joint frailty-copula models: PO_{EX} , PO_{PE} when the generator is equivalent to the PO_{EX} ($L = 300$ and $M_C = 250$).



Source: Prepared by the author.

and $\phi_1^{(T)}$ exhibited significant discrepancies, although the proportion of outliers did not exceed 3.6% of the total number of posterior samples.

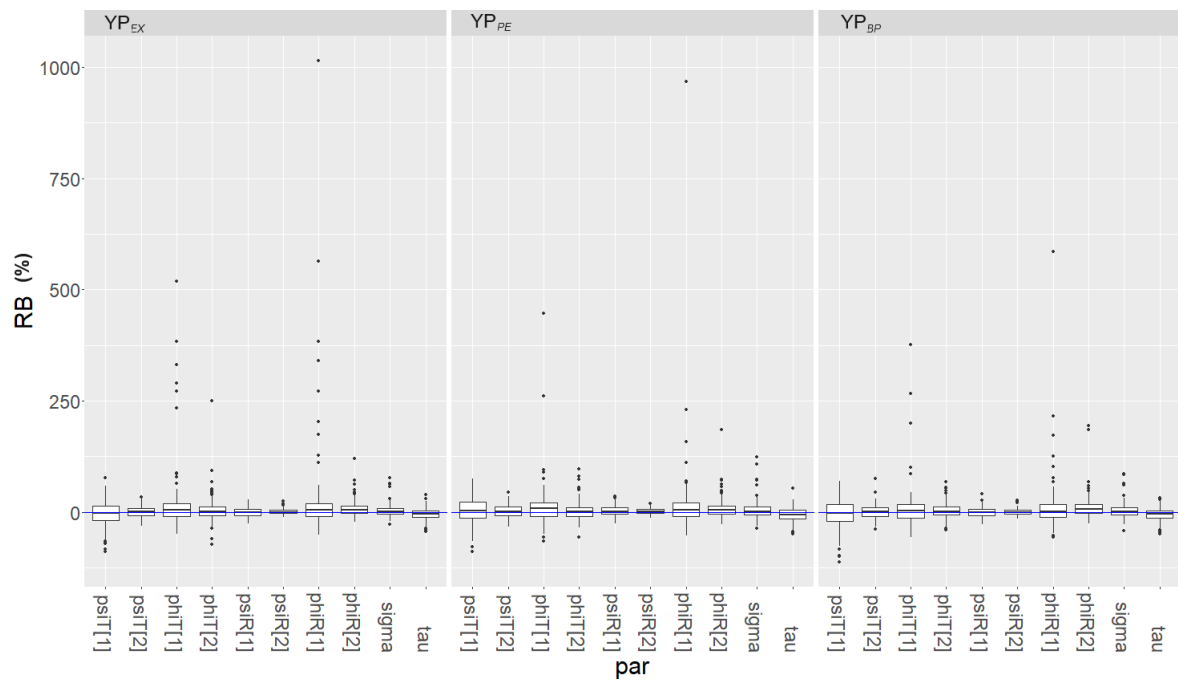
On the other hand, when considering the median biases, all parameters in the three models demonstrated values close to zero. Similar to the other regression families, the τ_κ exhibits negative medians, albeit close to zero. The variability of the biases for this parameter was nearly symmetrical, and the mean biases were also negative.

In terms of the long-term regression coefficients, their biases exhibited a more symmetrical behavior around the median, resulting in similar mean and median biases. As for the parameter σ_w , it generated right-skewed biases with a positive median close to zero.

These analyses provide an understanding of the biases within the YP regression family. Although there are some outliers and discrepancies in certain parameters, the majority of the biases, as reflected by the median values, are close to zero. This suggests that the YP models were generally able to capture the characteristics of data. All simulation study results shown in this section are available online²

²Access the link cassiushenrique.shinyapps.io/appSimulationsJointFrailtyCopula.

Figure 4.8: Boxplot of RB(%) for the joint frailty-copula models: YP_{EX} , YP_{PE} when the generator is equivalent to the YP_{EX} ($L = 300$ and $M_C = 250$)



Source: Prepared by the author.

Chapter 5

Data analysis

In this chapter, we present the application of our models. Works such as [Rondeau et al. \(2007\)](#) and [Li et al. \(2019\)](#) illustrate the application of survival models to individuals with recurrent events and dependent censoring caused by a terminal event. To illustrate the application of our proposal, we use a database called `readmission` from the `frailtypack` package ([Rondeau et al., 2012](#)), previously applied in the study of [González et al. \(2005\)](#). The data originates from Bellvitge's Public University Hospital in Barcelona, Spain, capturing records from January 1996 to December 1998. Out of 523 newly diagnosed colorectal cancer patients, that study focused on the 403 who underwent surgery. The study's response variable is readmission time (in days), considering it as a potentially recurrent event since patients with colorectal cancer may have several readmissions after discharge. The study began on each patient's surgery date, resulting in varying follow-up durations. Some premature follow-up termination occurred in cases of patient death, migration, or hospital transfer.

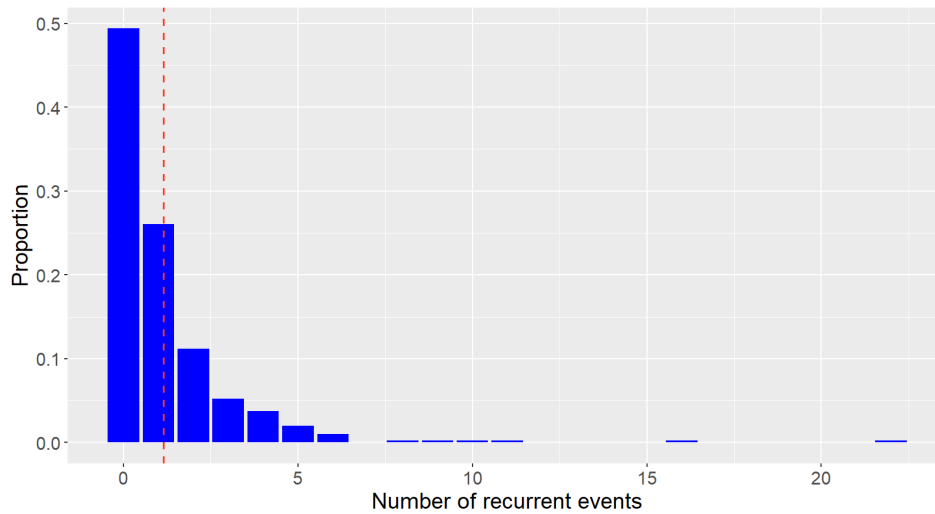
The first readmission time was considered as the interval between the date of the surgical procedure and the first readmission related to colorectal cancer. Subsequent readmission times were considered as the difference between the last discharge date and the current hospitalization date. This information was collected from the discharge diagnosis recorded by the clinical documentation department.

In the sample, descriptive statistics revealed of the 403 individuals, only 103 of them died, which means that administrative censorship reached a level of 72.95%. Furthermore, it was observed an average of approximately 1.14 readmissions per individual, with a standard deviation of about 2.02. While the highest observed number of readmissions for a single individual was 22, the median stood at one readmission, indicating a skewed distribution. Notably, a mere 10 subjects experienced over five readmissions each. The graph displayed in [Figure 5.1](#) shows these readmission frequencies. It also includes a red dashed line indicating the average number of recurrences.

From this dataset, four time-fixed effects recorded in the file will be considered:

- `sex` (Male, when `sex = 0`, or Female, when `sex=1`),
- `chemo` which represents whether there was chemotherapy treatment (`Treated`, when

Figure 5.1: Proportion of the number of readmissions.



Source: Prepared by the author.

`chemo=1`, or `nonTreated`, when `chemo=0`) and,

- `dukes` which represents the Dukes' stage. [González et al. \(2005\)](#) classified the sample patients as (A-B, C or D). These stages range from cancer within the inner intestinal layer to distant spread. The Dukes' stages are used to classify colorectal cancer into different stages based on the extent of the tumor into four main stages: Dukes A, when the cancer is in the inner layer of the intestine or growing slightly in the muscular layer; Dukes B, when cancer has grown through the muscular layer of the intestine; Dukes C, if cancer has spread to at least one lymph node near the intestine and, Dukes D, if cancer has spread to another part of the body, such as the liver, lungs or bones ([Wong et al., 2004](#)). Table 5.1 shows how we configure two dummy variables to accommodate the three levels of Dukes' stages, A-B, C, and D.

Table 5.1: Dummy variable for variable `dukes`.

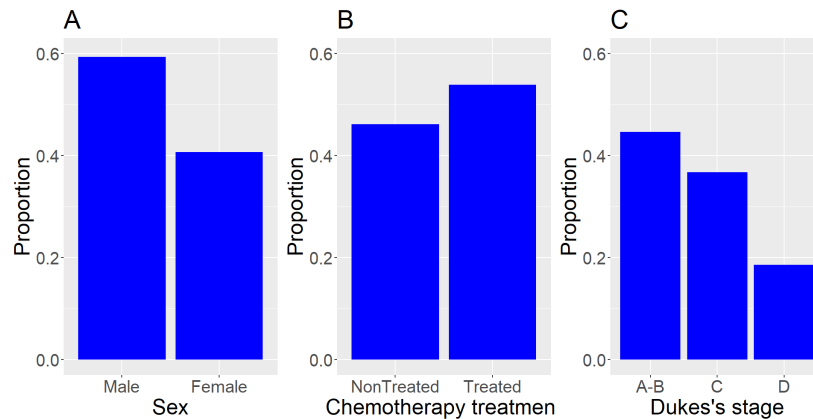
Dukes' stages	Dukes ₁	Dukes ₂
A-B	0	0
C	1	0
D	0	1

Another characteristic available in the `readmission` database is the Charlson index. Many individuals showed changes in the value of this index over time. Initially developed by [Charlson et al. \(1987\)](#), this index is widely used in medical research and clinical practice to adjust or control the influence of comorbidities in epidemiological studies and analyses of health outcomes. [González et al. \(2005\)](#) used an adaptation of this index proposed by [Librero et al. \(1999\)](#). As our class of models does not accommodate

time-dependent explanatory variables, it will be disregarded in this application. In future work, the models could be adapted to accommodate this type of covariate.

Figure 5.2 shows the proportion of the categories of the covariates. We observed that 164 were women, 239 were men; 217 received chemotherapy treatment and 186 did not. Colorectal cancer of 180 patients was classified as Dukes' stage A-B, of 148 as Dukes' stage C, and of 75 as Dukes' stage D.

Figure 5.2: Proportion of the categories of the covariates (A) sex, (B) chemotherapy treatment, and (C) Dukes's stage.



Source: Prepared by the author.

In this chapter, we will present the application in two sections: Section 5.1 provides an analysis of the models that are in our first class. In this case, we only focus on terminal events and use individual frailties to explain non-observed heterogeneities. Moving beyond the analysis of terminal events, Section 5.2 focuses on the influence of readmissions on survival time. In that section, we adopt a more holistic approach, by fitting our second class, considering not just terminal events but also recurrent events (hospital readmissions). This analysis is particularly interesting as it explores the interplay between terminal and recurrent events, offering a comprehensive view of patient trajectories.

To make our comparative analysis more robust, we use an objective criterion. The Widely Applicable Information Criterion (WAIC) is a statistic for model comparison, especially useful in Bayesian contexts (Ninomiya, 2021). It is a generalization of the well-known Akaike Information Criterion (AIC) and is applicable even when the model is complex or when the number of parameters is large concerning the number of observations (Akaike, 2011).

WAIC calculates the log-likelihood of the data given a model, denoted as $\widehat{\text{lppd}}$, and penalizes for the complexity of this model, considering the effective number of parameters. Lower WAIC values indicate a model with a better predictive fit. Unlike AIC, WAIC is based on a weighted average of all parameter posterior distributions rather than just a point estimate (Vehtari et al., 2023). Mathematically, it is an alternative approach to

estimating the expected log pointwise predictive density and is calculated by

$$\widehat{\text{elppd}}_{\text{waic}} = \widehat{\text{lppd}} - \widehat{p}_{\text{waic}},$$

where $\widehat{p}_{\text{waic}}$ is the estimated effective number of parameters and is computed based on the sum of the posterior variance of the log-likelihood function. In practical terms, we can calculate using the posterior variance of the log predictive density function for each data point y_i , i.e.,

$$\widehat{p}_{\text{waic}} = V_{q=1}^Q [\log p(y_i | \Theta^{(q)})],$$

in which

$$V_{q=1}^Q a_q = \frac{1}{Q} \sum_{q=1}^Q (a_q - \bar{a})^2.$$

Then, we define

$$\text{WAIC} = -2 \widehat{\text{elppd}}_{\text{waic}}. \quad (5.1)$$

To compare the fits of our models, we applied the function `waic` available in the package `loo` (Vehtari et al., 2021) which uses the previous expression to calculate the WAIC value based on each model.

We can use the R-hat to evaluate the convergence of the parameters more carefully. In MCMC methods, R-hat, also known as the Gelman-Rubin statistic, is a diagnostic tool used to assess the convergence of the MCMC algorithm (Gelman et al., 1995). It compares the variance between multiple chains to the variance within each chain. If the chains have converged to the target distribution, the between-chain, and within-chain variances should be similar, yielding an R-hat value close to 1. Values of R-hat substantially greater than 1 indicate that the chains may not have converged, suggesting that either more iterations are needed or the model requires meliorations (Gelman et al., 1995; Peng, 2020).

In all models whose baseline is piecewise exponential or Bernstein Polynomials, we use $m = 5$. In future work, we can perform a sensitivity analysis of the choice of m on the WAIC values.

5.1 Analysis of the Yang and Prentice frailty model (Class 1)

We initiate our application by fitting the first class of models. For the inference procedure, we generated four MCMC chains for each parameter via `rstan` (Stan Development Team, 2018) with 5000 iterations, of which 2500 are warm-ups, resulting in posterior

samples of size 10000. To reduce a possible autocorrelation effect of the σ_w parameter, we choose to record the posterior systematic samples with a period equal to 3. This way, of every three MCMC samples, only the last one was saved. This setting was adjusted by setting `thin = 3` in the `rstan::sampling` function. In this section, we show the posterior means of these samples, their standard deviation as well as their 95% credible intervals (CI).

We choose weakly informative prior as classified by [Stan Development Team \(2023\)](#). They are listed as follows:

$$\begin{aligned}\psi_1, \dots, \psi_p &\sim \text{Normal}(0, 3), \\ \phi_1, \dots, \phi_p &\sim \text{Normal}(0, 3), \\ \gamma_1, \dots, \gamma_m &\sim \text{LogNormal}(0, 2), \\ \sigma_w &\sim \text{Gamma}(1, 1),\end{aligned}$$

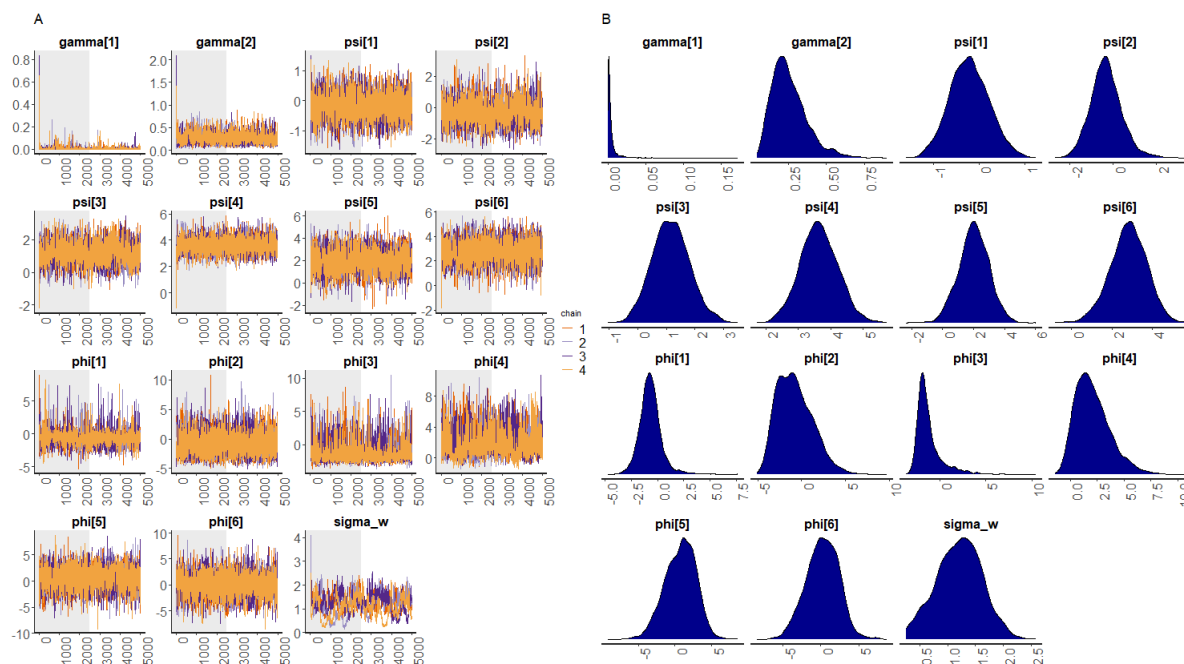
where m represents the dimensions of the baseline functions, while p denotes the covariate vector length, considering interactions between some covariates.

The WAIC estimates and the statistical summary provided by Tables 5.2 and 5.3 show the application from our first class of models that introduced a frailty term in the likelihood. We indicate significant variables by highlighting point estimates in bold. A significant covariate is defined as one whose credible interval does not contain zero. We show the 95% credible interval of each parameter, the standard deviation of its estimates, and the WAIC score of the models, highlighting in bold the one with the lowest WAIC. The primary interest is in the model with the best WAIC score, which is YP_{BP} . Therefore, let's explore the estimates found from this model. It is interesting to note that the PH_{BP} model was in second place, with a very close WAIC value.

In YP_{BP} , the parameters of the baseline hazard function $\boldsymbol{\gamma}$ are equivalents as φ in Section 2.4. They will not be discussed here, since they do not have a direct interpretation. The estimate for `sex` variable is negative, both in the short and long term. The negative value suggests that being female (`sex=1`) might be associated with a lower probability of death compared to being male (`sex=0`), but the credible interval includes 0, indicating this result is not statistically significant. The variable `chemo` is also not significant. The estimate for `Dukes2` is positive and significant, in the short and long terms. Its value indicates that being at Dukes' stage D in comparison to A-B is associated with a higher likelihood of death. The interaction between chemotherapy treatment and Dukes' stage is significant in the short term. This means that individuals with cancer in more advanced stages, even if they undergo chemotherapy at the beginning of follow-up, have a higher risk of death. We note that the estimates of the parameter σ_w suggest that there is variability among study individuals that is not explained by the model's fixed effects alone.

Figure 5.3-A shows MCMC graphs and Figure 5.3-B presents the posterior densities provided for YP_{BP} . The MCMC chains for the regression coefficients appear to be

Figure 5.3: MCMC applied to YP_{BP} model: (A) Trace plots for the posterior samples; (B) Posterior density plots.



Source: Prepared by the author.

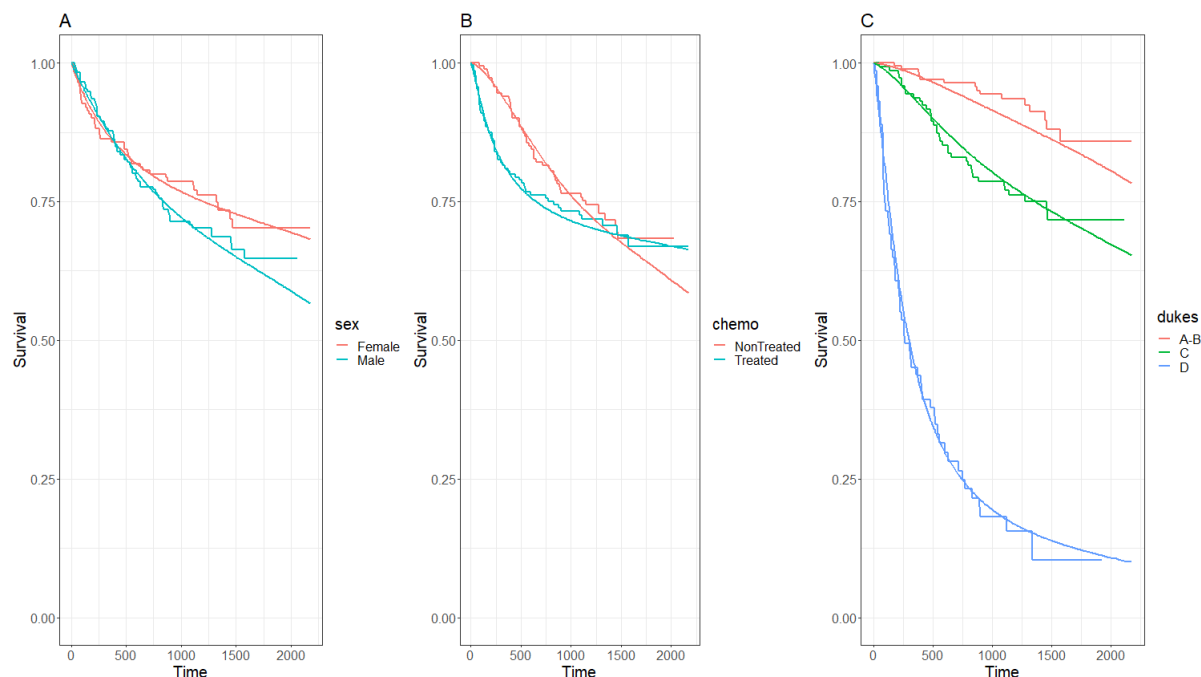
mixing well, with overlap between the four chains and no visible trends, which indicates convergence. The standard deviation of the frailty σ_w shows a reasonable convergence between the chains. The estimates of R-hat for the parameters are between 0.9995 and 1.1243 indicating that convergence occurred. Observing the posterior densities we note that they are all unimodal.

From Figure 5.4, we can notice a similarity between the survival functions estimated by our model (represented by the continuous curves) and the Kaplan-Meier estimates. This fact suggests that our model is capturing the pattern of the observed data. However, it is perceived the high volume of censoring in this dataset. Our models do not include cure fraction models. The inclusion of this approach could increase the precision of long-term effect estimates. It is anticipated that incorporating cure fraction models into our class of models will be done in future research.

Figure 5.4-A displays the survival curves for both men and women. This figure suggests that there is insufficient statistical evidence to conclude that a patient's sex significantly influences survival time. Figure 5.4-B presents the short-term and long-term impact of the chemotherapy treatment. The curves are further apart at the beginning of the follow-up and get closer at the end of the follow-up. However, individuals with cancer in more advanced stages, even those receiving chemotherapy, tend to be at greater risk of death. In Figure 5.4-C, we see the survival functions for Dukes' stages A-B, C, and D present well-defined differences between them. The Dukes' stage D shows the lowest survival functions over time, while patients with Dukes's stages A-B colorectal cancer

appear to survive longer.

Figure 5.4: Kaplan-Meier (step function) and survival curves estimated by YP_{BP} model (continuous function) about the terminal event for the levels of variables (A) **sex**, (B) **chemo**, and (C) **dukes**. Time is measured in days.



The estimates of the other models are found in Table 5.2 whose MCMC and posterior density graphs can be seen online¹. The PH and PO models agreed that the patient's sex is the only non-significant variable at 95% confidence. This classification was also observed in the short-term effects of the YP models.

5.2 Analysis of the joint frailty-copula models (Class 2)

After analyzing our model in the context of terminal events, we expanded our focus to include both terminal and recurrent events. For this more comprehensive evaluation, the implementation of our joint frailty-copula models will be useful. These models are in the second class. They also allow us to assess the strength of the association between hospital readmissions and deaths.

For the inference procedure, we use the default of `rstan::sampling` function. It means that we generate four MCMC chains for each parameter using the `rstan` package

¹Access the link cassiushenrique.shinyapps.io/appRealFrailty.

Table 5.2: Summary of the PH and PO models fitted to the readmission data considering the terminal events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.

model	WAIC	par	description	est	sd	95% CI			
						LW	UP		
PH _{EX}	1831.34	β_1	Sex	-0.2528	0.2150	-0.6833	0.1566		
		β_2	Chemo	-0.4221	0.4976	-1.3829	0.5902		
		β_3	Dukes ₁	0.6181	0.4413	-0.2241	1.5359		
		β_4	Dukes ₂	2.6718	0.4580	1.8333	3.6204		
		β_5	Chemo * Dukes ₁	1.2072	0.6057	0.0044	2.4112		
		β_6	Chemo * Dukes ₂	1.5377	0.5856	0.3664	2.6595		
		σ_w	sd(Frailty)	0.5445	0.2521	0.1181	1.0001		
		γ_1	baseline	0.2014	0.0796	0.0765	0.3903		
		PH _{PE}	1824.28	β_1	Sex	-0.3038	0.2356	-0.7899	0.1312
β_2	Chemo			-0.4841	0.4972	-1.4680	0.4893		
β_3	Dukes ₁			0.5737	0.4465	-0.2636	1.4575		
β_4	Dukes ₂			2.7978	0.4992	1.8577	3.8196		
β_5	Chemo * Dukes ₁			1.3361	0.6228	0.0951	2.5346		
β_6	Chemo * Dukes ₂			1.7531	0.6123	0.5562	2.9583		
σ_w	sd(Frailty)			0.8243	0.2152	0.4304	1.2695		
PH _{BP}	1798.27			β_1	Sex	-0.5138	0.3357	-1.2268	0.0904
				β_2	Chemo	-0.7478	0.5862	-1.9153	0.3705
		β_3	Dukes ₁	0.5368	0.5265	-0.4980	1.5807		
		β_4	Dukes ₂	3.6270	0.7175	2.2175	5.0647		
		β_5	Chemo * Dukes ₁	1.9482	0.8147	0.4213	3.6343		
		β_6	Chemo * Dukes ₂	2.7450	0.9205	1.0549	4.5723		
		σ_w	sd(Frailty)	1.6462	0.4532	0.6968	2.3694		
		PO _{EX}	1834.81	β_1	Sex	-0.3343	0.2565	-0.8486	0.1500
				β_2	Chemo	-0.4872	0.4996	-1.4413	0.5235
β_3	Dukes ₁			0.5854	0.4530	-0.2569	1.5322		
β_4	Dukes ₂			2.7849	0.4786	1.8860	3.7793		
β_5	Chemo * Dukes ₁			1.3830	0.6455	0.0852	2.6272		
β_6	Chemo * Dukes ₂			1.8384	0.6185	0.6295	3.0458		
σ_w	sd(Frailty)			0.2263	0.1613	0.0300	0.5997		
γ_1	baseline			0.2464	0.0889	0.1037	0.4477		
PO _{PE}	1833.48			β_1	Sex	-0.3452	0.2540	-0.8484	0.1530
		β_2	Chemo	-0.4673	0.5408	-1.5358	0.5985		
		β_3	Dukes ₁	0.6367	0.4773	-0.2992	1.6272		
		β_4	Dukes ₂	3.0335	0.5299	2.0176	4.0840		
		β_5	Chemo * Dukes ₁	1.4364	0.6715	0.0839	2.7385		
		β_6	Chemo * Dukes ₂	1.9408	0.6933	0.6341	3.3162		
		σ_w	sd(Frailty)	0.2484	0.2112	0.0034	0.7480		
		PO _{BP}	1819.15	β_1	Sex	-0.5107	0.3328	-1.1926	0.0980
				β_2	Chemo	-0.5407	0.5847	-1.7126	0.5632
β_3	Dukes ₁			0.7338	0.5149	-0.2934	1.7539		
β_4	Dukes ₂			3.8928	0.6548	2.6767	5.2011		
β_5	Chemo * Dukes ₁			1.7880	0.8148	0.2876	3.4728		
β_6	Chemo * Dukes ₂			2.6787	0.8903	1.0115	4.4661		
σ_w	sd(Frailty)			1.2280	0.5505	0.1069	2.0402		

(Stan Development Team, 2018), each comprising 2000 iterations, of which the initial 1000 iterations are of warm-up. Consequently, this process yielded posterior samples with a total size of 1000. We evaluate our models in terms of their average posterior estimate, the posterior standard deviation, and their respective 95% credible intervals (CI).

We have opted for weakly informative priors, as categorized by Stan Development

Table 5.3: Summary of the YP models fitted to the readmission data considering the terminal events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.

model	WAIC	par	description	est	sd	95% CI			
						LW	UP		
YP _{EX}	1832.35	ψ_1	Sex	-0.2717	0.2565	-0.7713	0.2511		
		ψ_2	Chemo	-0.4269	0.5388	-1.4898	0.6541		
		ψ_3	Dukes ₁	0.6157	0.4876	-0.3009	1.6039		
		ψ_4	Dukes ₂	2.6540	0.4854	1.7134	3.6474		
		ψ_5	Chemo * Dukes ₁	1.2468	0.6817	-0.0870	2.5966		
		ψ_6	Chemo * Dukes ₂	1.6725	0.6586	0.3847	2.9726		
		ϕ_1	Sex	0.9042	2.1382	-2.5312	5.8734		
		ϕ_2	Chemo	1.1207	2.2117	-2.5051	5.9694		
		ϕ_3	Dukes ₁	1.8497	2.0123	-1.1997	6.4772		
		ϕ_4	Dukes ₂	2.9056	1.6995	0.1847	6.8180		
		ϕ_5	Chemo * Dukes ₁	0.3714	2.6935	-4.5275	6.0358		
		ϕ_6	Chemo * Dukes ₂	-0.1321	2.6694	-4.9818	5.1890		
		σ_w	sd(Frailty)	0.5310	0.2328	0.1597	0.9950		
		γ_1	baseline	0.2055	0.0825	0.0823	0.3952		
		YP _{PE}	1830.92	ψ_1	Sex	-0.2731	0.2857	-0.8162	0.3125
				ψ_2	Chemo	-0.4393	0.5390	-1.4933	0.6224
				ψ_3	Dukes ₁	0.6341	0.4917	-0.3066	1.6425
ψ_4	Dukes ₂			2.7646	0.4904	1.8464	3.7775		
ψ_5	Chemo * Dukes ₁			1.3920	0.7154	0.0586	2.8562		
ψ_6	Chemo * Dukes ₂			1.8363	0.6664	0.5550	3.1525		
ϕ_1	Sex			0.5027	1.9652	-2.5778	5.3000		
ϕ_2	Chemo			0.6188	2.1787	-2.8811	5.3461		
ϕ_3	Dukes ₁			1.3993	2.0395	-1.6015	6.0357		
ϕ_4	Dukes ₂			2.7956	1.8234	0.0324	7.0962		
ϕ_5	Chemo * Dukes ₁			0.3629	2.7057	-4.5682	6.2907		
ϕ_6	Chemo * Dukes ₂			-0.1198	2.5105	-4.9650	5.1718		
σ_w	sd(Frailty)			0.6235	0.2933	0.1516	1.2229		
YP _{BP}	1794.64			ψ_1	Sex	-0.2122	0.4257	-1.0333	0.6558
				ψ_2	Chemo	-0.3307	0.7277	-1.6513	1.2817
				ψ_3	Dukes ₁	1.1541	0.6554	-0.1094	2.5243
				ψ_4	Dukes ₂	3.6413	0.5822	2.5000	4.8180
		ψ_5	Chemo * Dukes ₁	2.0182	0.9840	0.0650	3.9126		
		ψ_6	Chemo * Dukes ₂	2.8088	0.9107	0.9140	4.5319		
		ϕ_1	Sex	-0.6734	1.1280	-2.6184	2.0198		
		ϕ_2	Chemo	-0.7680	2.0203	-3.8750	3.6330		
		ϕ_3	Dukes ₁	-1.1885	1.4296	-2.7703	2.8973		
		ϕ_4	Dukes ₂	2.0736	1.6249	-0.3909	5.8904		
		ϕ_5	Chemo * Dukes ₁	0.7192	2.1300	-3.6591	4.4600		
		ϕ_6	Chemo * Dukes ₂	0.2049	2.1791	-4.1632	4.1804		
		σ_w	sd(Frailty)	1.2043	0.4062	0.3946	1.9889		

Team (2023), which are outlined as follows:

$$\begin{aligned} \theta &\sim \text{Gamma}(1, 1), \\ \sigma_w &\sim \text{Gamma}(1, 1), \\ \gamma_1^{(R)}, \dots, \gamma_{m_R}^{(R)} &\sim \text{LogNormal}(0, 2), \\ \gamma_1^{(T)}, \dots, \gamma_{m_T}^{(T)} &\sim \text{LogNormal}(0, 2), \\ \psi_1^{(R)}, \dots, \psi_p^{(R)} &\sim \text{Normal}(0, 3), \\ \psi_1^{(T)}, \dots, \psi_p^{(T)} &\sim \text{Normal}(0, 3), \\ \phi_1^{(R)}, \dots, \phi_p^{(R)} &\sim \text{Normal}(0, 3), \end{aligned}$$

and

$$\phi_1^{(T)}, \dots, \phi_p^{(T)} \sim \text{Normal}(0, 3),$$

where m_R and m_T represent the dimensions of the baseline functions for recurrent and terminal events, respectively, while p denotes the dimension of the linear predictor vectors.

The estimates derived from the PH models are presented in Table 5.4, and from the PO models in Table 5.5. Tables 5.6, 5.7 and 5.8 provide the estimates from the YP models. Additionally, these tables include the WAIC values for each respective model. Notably, the YP_{BP} demonstrates the most favorable WAIC, indicating its superior performance in comparison to the others. Therefore, in this section, we will focus on it. As the parameters of the baseline hazard function are not interpretable, we will not display its results here.

The variable **sex** was identified as significant only in the short-term effects on readmissions. Its negative coefficient suggests that females (**sex**=1) have a lower risk of hospital readmissions compared to males (**sex**=0). Chemotherapy treatment is not significant in both short-term and long-term effects on death, because the credible intervals contain zero. In contrast, the treatment is significant on readmissions. The estimates of the regression coefficient of this variable indicate a crossover in the survival curves concerning readmissions. At the beginning of follow-up, patients undergoing chemotherapy faced a higher readmission risk compared to those who did not receive treatment. However, in the long term, the opposite trend was observed. However, we identified that the interaction between Dukes' stage and chemotherapy is significant in readmissions. Patients whose cancer is in more advanced stages, even receiving chemotherapy, are at greater risk of experiencing new cancer-related hospital readmissions at the end of the follow-up. Dukes' stage D is significant in death risk in the short term, however, in the long term, the Dukes' stages are not significant. Additionally, in the short term, the Dukes' stage of cancer is significant in the readmissions.

The estimates of the standard deviation of frailty lead us to conclude that there is some association between the readmissions of the patients. The interval estimates of the Kendall correlation coefficient indicate a small correlation between readmissions and death in the investigated colorectal cancer patients.

Figure 5.5-A displays the MCMC graphs for the YP_{BP}. The MCMC chains for the coefficients are well-mixing, evidenced by the overlap among the four chains. The R-hat estimates for the parameters range from 0.9994 to 1.0020. Furthermore, in Figure 5.5-B we see the posterior densities are unimodal. The MCMC and posterior density graphs for the other models are available online².

²Access the link cassiushenrique.shinyapps.io/appRealJointFrailtyCopula.

Figure 5.5: MCMC applied to YP_{BP} model: (A) posterior trace plots for the posterior samples; (B) posterior density plots.

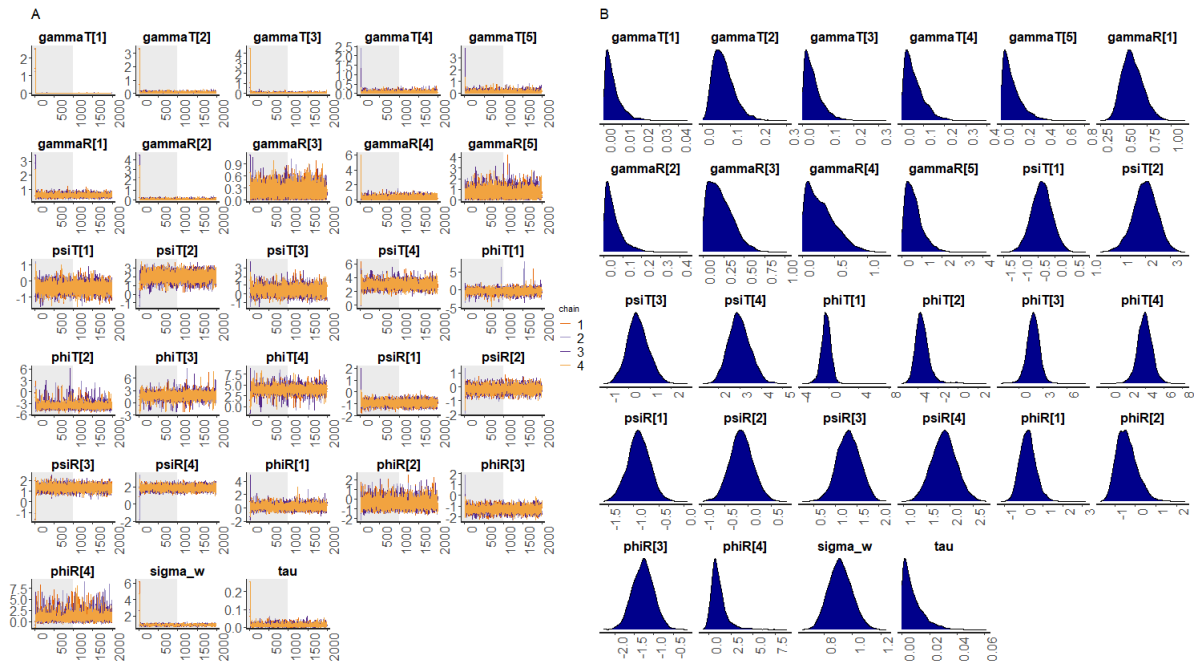


Table 5.4: Summary of the PH models fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.

model	WAIC	par	description	est	sd	95% CI			
						LW	UP		
PH _{EX}	2282.12	$\gamma_1^{(T)}$	baseline	0.1594	0.0669	0.0594	0.3159		
		$\gamma_1^{(R)}$	baseline	1.3974	0.3292	0.8479	2.1126		
		$\beta_1^{(T)}$	Sex	-0.2618	0.2476	-0.7448	0.2115		
		$\beta_2^{(T)}$	Chemo	-0.4937	0.5294	-1.5355	0.5806		
		$\beta_3^{(T)}$	Dukes ₁	0.5364	0.4776	-0.3650	1.4922		
		$\beta_4^{(T)}$	Dukes ₂	2.6943	0.5154	1.7143	3.7023		
		$\beta_5^{(T)}$	Chemo * Dukes ₁	1.3723	0.6665	0.0662	2.6797		
		$\beta_6^{(T)}$	Chemo * Dukes ₂	1.8519	0.6485	0.5258	3.1114		
		$\beta_1^{(R)}$	Sex	-0.5558	0.1712	-0.8920	-0.2301		
		$\beta_2^{(R)}$	Chemo	-0.2919	0.2805	-0.8635	0.2484		
		$\beta_3^{(R)}$	Dukes ₁	0.3199	0.2745	-0.2012	0.8603		
		$\beta_4^{(R)}$	Dukes ₂	2.0169	0.3309	1.3794	2.6581		
		$\beta_5^{(R)}$	Chemo * Dukes ₁	0.4429	0.4093	-0.3561	1.2505		
		$\beta_6^{(R)}$	Chemo * Dukes ₂	0.2593	0.4359	-0.6044	1.1356		
		σ_w	sd(Frailty)	1.1818	0.0856	1.0195	1.3539		
		τ_κ	Kendall's tau	0.0072	0.0064	0.0002	0.0236		
		PH _{PE}	2261.98	$\beta_1^{(T)}$	Sex	-0.2775	0.2305	-0.7373	0.1524
				$\beta_2^{(T)}$	Chemo	-0.6994	0.4734	-1.6249	0.2236
$\beta_3^{(T)}$	Dukes ₁			0.3278	0.4158	-0.4723	1.1787		
$\beta_4^{(T)}$	Dukes ₂			2.4460	0.4387	1.5726	3.3219		
$\beta_5^{(T)}$	Chemo * Dukes ₁			1.5025	0.6135	0.3131	2.7261		
$\beta_6^{(T)}$	Chemo * Dukes ₂			2.1059	0.5958	0.9746	3.2937		
$\beta_1^{(R)}$	Sex			-0.4804	0.1513	-0.7735	-0.1904		
$\beta_2^{(R)}$	Chemo			-0.2589	0.2411	-0.7279	0.2238		
$\beta_3^{(R)}$	Dukes ₁			0.2868	0.2361	-0.1643	0.7491		
$\beta_4^{(R)}$	Dukes ₂			1.5316	0.2885	0.9817	2.1061		
$\beta_5^{(R)}$	Chemo * Dukes ₁			0.2990	0.3497	-0.3848	0.9850		
$\beta_6^{(R)}$	Chemo * Dukes ₂			-0.0034	0.3869	-0.7461	0.7543		
σ_w	sd(Frailty)			0.9085	0.0804	0.7556	1.0725		
τ_κ	Kendall's tau			0.0083	0.0080	0.0002	0.0300		
PH _{BP}	2262.59			$\beta_1^{(T)}$	Sex	-0.3328	0.2333	-0.8176	0.1131
				$\beta_2^{(T)}$	Chemo	-0.9931	0.4718	-1.9200	-0.0710
				$\beta_3^{(T)}$	Dukes ₁	0.0341	0.4083	-0.7325	0.8476
				$\beta_4^{(T)}$	Dukes ₂	2.2431	0.4307	1.4140	3.0979
		$\beta_5^{(T)}$	Chemo * Dukes ₁	1.8201	0.6153	0.6223	2.9995		
		$\beta_6^{(T)}$	Chemo * Dukes ₂	2.5683	0.6002	1.4345	3.7479		
		$\beta_1^{(R)}$	Sex	-0.5302	0.1541	-0.8446	-0.2335		
		$\beta_2^{(R)}$	Chemo	-0.4336	0.2511	-0.9238	0.0561		
		$\beta_3^{(R)}$	Dukes ₁	0.1391	0.2471	-0.3384	0.6268		
		$\beta_4^{(R)}$	Dukes ₂	1.4984	0.2932	0.9460	2.0607		
		$\beta_5^{(R)}$	Chemo * Dukes ₁	0.4865	0.3694	-0.2105	1.2284		
		$\beta_6^{(R)}$	Chemo * Dukes ₂	0.2872	0.3987	-0.4764	1.0790		
		σ_w	sd(Frailty)	0.9891	0.0839	0.8356	1.1616		
		τ_κ	Kendall's tau	0.0077	0.0073	0.0002	0.0274		

Table 5.5: Summary of the PO models fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.

model	WAIC	par	description	est	sd	95% CI			
						LW	UP		
PO _{EX}	2342.48	$\gamma_1^{(T)}$	baseline	0.1058	0.0522	0.0353	0.2364		
		$\gamma_1^{(R)}$	baseline	0.9202	0.1803	0.6085	1.3138		
		$\beta_1^{(T)}$	Sex	-0.3198	0.3161	-0.9573	0.2831		
		$\beta_2^{(T)}$	Chemo	-0.4462	0.6167	-1.6652	0.7408		
		$\beta_3^{(T)}$	Dukes ₁	0.8656	0.5695	-0.2592	1.9714		
		$\beta_4^{(T)}$	Dukes ₂	3.5241	0.6320	2.3117	4.7782		
		$\beta_5^{(T)}$	Chemo * Dukes ₁	1.5257	0.7970	0.0071	3.0865		
		$\beta_6^{(T)}$	Chemo * Dukes ₂	1.8387	0.8086	0.2542	3.4522		
		$\beta_1^{(R)}$	Sex	-0.6555	0.2305	-1.1119	-0.2137		
		$\beta_2^{(R)}$	Chemo	0.3294	0.3236	-0.3232	0.9463		
		$\beta_3^{(R)}$	Dukes ₁	1.2548	0.3241	0.6146	1.8805		
		$\beta_4^{(R)}$	Dukes ₂	3.0076	0.3962	2.2345	3.7873		
		$\beta_5^{(R)}$	Chemo * Dukes ₁	-0.3761	0.5194	-1.3628	0.6588		
		$\beta_6^{(R)}$	Chemo * Dukes ₂	-0.6677	0.5818	-1.7976	0.5188		
		σ_w	sd(Frailty)	1.5362	0.1128	1.3247	1.7650		
		τ_κ	Kendall's tau	0.0105	0.0096	0.0003	0.0359		
		PO _{PE}	2292.66	$\beta_1^{(T)}$	Sex	-0.4109	0.3146	-1.0271	0.1929
				$\beta_2^{(T)}$	Chemo	-0.8261	0.5500	-1.8943	0.2479
$\beta_3^{(T)}$	Dukes ₁			0.4139	0.4986	-0.5432	1.4149		
$\beta_4^{(T)}$	Dukes ₂			3.2861	0.5841	2.1145	4.4280		
$\beta_5^{(T)}$	Chemo * Dukes ₁			1.9799	0.7337	0.6054	3.4124		
$\beta_6^{(T)}$	Chemo * Dukes ₂			2.8436	0.7735	1.3537	4.3445		
$\beta_1^{(R)}$	Sex			-0.6990	0.2169	-1.1249	-0.2872		
$\beta_2^{(R)}$	Chemo			-0.1858	0.3183	-0.8469	0.4086		
$\beta_3^{(R)}$	Dukes ₁			0.6466	0.3212	0.0220	1.2649		
$\beta_4^{(R)}$	Dukes ₂			2.1636	0.4111	1.3758	2.9685		
$\beta_5^{(R)}$	Chemo * Dukes ₁			0.0776	0.4864	-0.8419	1.0528		
$\beta_6^{(R)}$	Chemo * Dukes ₂			-0.2591	0.5484	-1.3279	0.8245		
σ_w	sd(Frailty)			1.3128	0.1173	1.0878	1.5504		
τ_κ	Kendall's tau			0.0087	0.0082	0.0003	0.0307		
PO _{BP}	2286.16			$\beta_1^{(T)}$	Sex	-0.4878	0.3139	-1.1054	0.1023
				$\beta_2^{(T)}$	Chemo	-1.1338	0.5465	-2.2028	-0.0872
				$\beta_3^{(T)}$	Dukes ₁	0.1211	0.4972	-0.8401	1.0978
				$\beta_4^{(T)}$	Dukes ₂	3.1230	0.5591	2.0085	4.2290
		$\beta_5^{(T)}$	Chemo * Dukes ₁	2.3357	0.7399	0.9106	3.7773		
		$\beta_6^{(T)}$	Chemo * Dukes ₂	3.3505	0.7711	1.8597	4.9019		
		$\beta_1^{(R)}$	Sex	-0.7505	0.2204	-1.1878	-0.3220		
		$\beta_2^{(R)}$	Chemo	-0.3072	0.3374	-0.9674	0.3466		
		$\beta_3^{(R)}$	Dukes ₁	0.5467	0.3270	-0.0972	1.1738		
		$\beta_4^{(R)}$	Dukes ₂	2.1231	0.4080	1.3056	2.9244		
		$\beta_5^{(R)}$	Chemo * Dukes ₁	0.2004	0.5067	-0.7844	1.2025		
		$\beta_6^{(R)}$	Chemo * Dukes ₂	-0.0896	0.5707	-1.2003	1.0314		
		σ_w	sd(Frailty)	1.3788	0.1170	1.1546	1.6120		
		τ_κ	Kendall's tau	0.0088	0.0083	0.0003	0.0311		

Table 5.6: Summary of the YP_{EX} model fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.

model	WAIC	par	description	est	sd	95% CI	
						LW	UP
YP_{EX}	2261.64	$\gamma_1^{(T)}$	baseline	0.2182	0.0888	0.0863	0.4331
		$\gamma_1^{(R)}$	baseline	2.2016	0.4160	1.4620	3.1036
		$\psi_1^{(T)}$	Sex	-0.3321	0.2506	-0.8174	0.1717
		$\psi_2^{(T)}$	Chemo	-0.5923	0.5423	-1.6593	0.4562
		$\psi_3^{(T)}$	Dukes ₁	0.4561	0.4893	-0.4843	1.4502
		$\psi_4^{(T)}$	Dukes ₂	2.5130	0.4915	1.5622	3.4857
		$\psi_5^{(T)}$	Chemo * Dukes[1]	1.3965	0.6670	0.0802	2.7037
		$\psi_6^{(T)}$	Chemo * Dukes[2]	1.8881	0.6531	0.5810	3.1382
		$\phi_1^{(T)}$	Sex	0.9909	2.2059	-2.6678	6.0189
		$\phi_2^{(T)}$	Chemo	1.4688	2.2345	-2.3834	6.2493
		$\phi_3^{(T)}$	Dukes ₁	1.9666	2.0351	-1.1484	6.5623
		$\phi_4^{(T)}$	Dukes ₂	3.2635	1.7416	0.4937	7.3547
		$\phi_5^{(T)}$	Chemo * Dukes[1]	0.3176	2.7229	-4.8826	6.0018
		$\phi_6^{(T)}$	Chemo * Dukes[2]	0.3357	2.6766	-4.6503	5.7090
		$\psi_1^{(R)}$	Sex	-1.2882	0.2471	-1.7698	-0.7997
		$\psi_2^{(R)}$	Chemo	1.4212	0.3546	0.7398	2.1128
		$\psi_3^{(R)}$	Dukes ₁	2.2120	0.3063	1.6075	2.8154
		$\psi_4^{(R)}$	Dukes ₂	2.5978	0.3383	1.9281	3.2488
		$\psi_5^{(R)}$	Chemo * Dukes[1]	-2.3454	0.5724	-3.4740	-1.1911
		$\psi_6^{(R)}$	Chemo * Dukes[2]	-1.9262	0.5529	-3.0173	-0.8213
		$\phi_1^{(R)}$	Sex	0.2328	0.2479	-0.2269	0.7308
		$\phi_2^{(R)}$	Chemo	-2.2666	0.2066	-2.6706	-1.8574
		$\phi_3^{(R)}$	Dukes ₁	-1.9170	0.1993	-2.3040	-1.5188
		$\phi_4^{(R)}$	Dukes ₂	-0.1799	0.3723	-0.8326	0.6261
		$\phi_5^{(R)}$	Chemo * Dukes[1]	2.8925	0.4486	2.0468	3.7829
		$\phi_6^{(R)}$	Chemo * Dukes[2]	2.5848	0.8868	1.1839	4.4222
		σ_w	sd(Frailty)	0.8582	0.0775	0.7129	1.0165
		τ_κ	Kendall's tau	0.0100	0.0096	0.0003	0.0369

Table 5.7: Summary of the YP_{PE} model fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.

model	WAIC	par	description	est	sd	95% CI	
						LW	UP
YP_{PE}	2255.11	$\psi_1^{(T)}$	Sex	-0.4129	0.2943	-0.9692	0.1973
		$\psi_2^{(T)}$	Chemo	-0.6129	0.5431	-1.6603	0.5115
		$\psi_3^{(T)}$	Dukes ₁	0.3887	0.4653	-0.4944	1.3338
		$\psi_4^{(T)}$	Dukes ₂	2.5694	0.4895	1.6296	3.5787
		$\psi_5^{(T)}$	Chemo * Dukes[1]	1.6753	0.7417	0.2498	3.2078
		$\psi_6^{(T)}$	Chemo * Dukes[2]	2.7069	0.7684	1.2490	4.2415
		$\phi_1^{(T)}$	Sex	0.5805	1.6251	-2.0940	4.7130
		$\phi_2^{(T)}$	Chemo	-0.0953	2.0165	-3.2136	4.5006
		$\phi_3^{(T)}$	Dukes ₁	1.3279	2.0528	-1.7578	5.9499
		$\phi_4^{(T)}$	Dukes ₂	2.3699	1.5605	-0.0347	5.8939
		$\phi_5^{(T)}$	Chemo * Dukes[1]	0.1220	2.4823	-4.6490	5.3225
		$\phi_6^{(T)}$	Chemo * Dukes[2]	-0.9888	2.3331	-5.3203	3.9500
		$\psi_1^{(R)}$	Sex	-1.0507	0.2522	-1.5549	-0.5646
		$\psi_2^{(R)}$	Chemo	0.9183	0.4074	0.1374	1.7050
		$\psi_3^{(R)}$	Dukes ₁	1.6601	0.3985	0.8897	2.4481
		$\psi_4^{(R)}$	Dukes ₂	2.0604	0.4089	1.2815	2.9012
		$\psi_5^{(R)}$	Chemo * Dukes[1]	-1.6953	0.6124	-2.8763	-0.4703
		$\psi_6^{(R)}$	Chemo * Dukes[2]	-1.3045	0.5854	-2.4984	-0.1908
		$\phi_1^{(R)}$	Sex	0.2886	0.3116	-0.2792	0.9269
		$\phi_2^{(R)}$	Chemo	-1.9590	0.3028	-2.5142	-1.2965
		$\phi_3^{(R)}$	Dukes ₁	-1.6378	0.2716	-2.1330	-1.0669
		$\phi_4^{(R)}$	Dukes ₂	0.9966	1.3406	-0.4491	4.7710
		$\phi_5^{(R)}$	Chemo * Dukes[1]	2.8414	0.6644	1.7339	4.2710
		$\phi_6^{(R)}$	Chemo * Dukes[2]	2.5495	1.7478	-0.8350	6.5352
σ_w		sd(Frailty)	0.8834	0.0848	0.7209	1.0519	
τ_κ		Kendall's tau	0.0094	0.0089	0.0003	0.0324	

Table 5.8: Summary of the YP_{BP} model fitted to the readmission data considering terminal and recurrent events: posterior mean estimate (est), standard deviation (sd) along with the 95% credible interval (LW, UP), and WAIC.

model	WAIC	par	description	est	sd	95% CI	
						LW	UP
YP_{BP}	2244.92	$\psi_1^{(T)}$	Sex	-0.5262	0.3318	-1.1591	0.1804
		$\psi_2^{(T)}$	Chemo	-0.7881	0.6057	-1.8527	0.5471
		$\psi_3^{(T)}$	Dukes ₁	0.2148	0.5134	-0.6928	1.3275
		$\psi_4^{(T)}$	Dukes ₂	2.4979	0.5088	1.5529	3.5745
		$\psi_5^{(T)}$	Chemo * Dukes[1]	2.2203	0.8457	0.5911	3.8469
		$\psi_6^{(T)}$	Chemo * Dukes[2]	3.4574	0.8035	1.8488	4.9818
		$\phi_1^{(T)}$	Sex	0.2239	0.9641	-1.4668	2.1291
		$\phi_2^{(T)}$	Chemo	-0.9719	1.7046	-3.6407	2.8316
		$\phi_3^{(T)}$	Dukes ₁	0.3260	1.8586	-2.2756	4.7853
		$\phi_4^{(T)}$	Dukes ₂	1.7294	1.4400	-0.4654	5.1862
		$\phi_5^{(T)}$	Chemo * Dukes[1]	-0.1006	2.2082	-4.4366	4.2702
		$\phi_6^{(T)}$	Chemo * Dukes[2]	-0.7985	1.9959	-4.9738	2.8192
		$\psi_1^{(R)}$	Sex	-1.2495	0.2520	-1.7337	-0.7498
		$\psi_2^{(R)}$	Chemo	1.0885	0.3924	0.3533	1.8768
		$\psi_3^{(R)}$	Dukes ₁	1.8709	0.3652	1.1826	2.6190
		$\psi_4^{(R)}$	Dukes ₂	2.2535	0.3868	1.5054	3.0304
		$\psi_5^{(R)}$	Chemo * Dukes[1]	-1.9697	0.5715	-3.0683	-0.8586
		$\psi_6^{(R)}$	Chemo * Dukes[2]	-1.5366	0.5660	-2.6667	-0.4461
		$\phi_1^{(R)}$	Sex	0.2429	0.2827	-0.3046	0.8144
		$\phi_2^{(R)}$	Chemo	-2.1654	0.2378	-2.6259	-1.6876
		$\phi_3^{(R)}$	Dukes ₁	-1.8396	0.2172	-2.2603	-1.4183
		$\phi_4^{(R)}$	Dukes ₂	0.1409	0.5257	-0.6639	1.3534
		$\phi_5^{(R)}$	Chemo * Dukes[1]	2.8638	0.4767	1.9577	3.8248
		$\phi_6^{(R)}$	Chemo * Dukes[2]	2.5603	1.1403	0.7224	5.2623
σ_w		sd(Frailty)	0.8801	0.0824	0.7194	1.0419	
τ_κ		Kendall's tau	0.0099	0.0090	0.0003	0.0335	

Chapter 6

Final remarks and future research

This thesis proposed to develop two classes of models within a Bayesian framework, designed to explain the impact of observed characteristics on survival curves that may intersect. For this finally, we used the YP regression structure for its ability to encompass and generalize the PH and PO models.

The first class of models embraces YP frailty. The incorporation of frailty in these models constitutes a contribution of this study, since in the literature the YP models did not incorporate frailty. We combined exponential, piecewise exponential, and Bernstein polynomials baseline functions. The selection of these last two baseline functions was motivated by their versatility because can fit a wide variety of hazard function shapes. In that regard, the innovations promoted by this thesis are the YP_{EX} , YP_{PE} , YP_{BP} , and PO_{EX} frailty models.

The models of the first class enable the analysis of survival data under distinct scenarios:

- individuals with a unique survival time where individual frailty explains unobserved heterogeneities;
- individuals organized in clusters for which the shared frailty explains a likely dependency in their survival times. This is because individuals from the same group may present certain similarities among them that are not observed when we compare them with individuals outside their group. Another way to apply shared frailty would be to evaluate the survival times of individuals who present recurrent events. In this case, the shared frailty accommodates the association between the survival times of the same individual. We can consider the individual as a cluster, thus, the frailty assumes a statistical context corresponding to when individuals are arranged in clusters.

The second class of models is another contribution of this thesis. This class embraces the joint frailty-copula models with three distinct families of regression: PH, PO, and YP. In each regression framework, three baseline functions also are considered: exponential, piecewise exponential, and Bernstein polynomials. These models are designed to model survival data involving individuals who may experience both recurrent events

and a terminal event. Within these models, the recurrent events attributed to an individual are linked by a frailty term, aimed at capturing potential associations among the times between recurrences. Furthermore, the terminal events are potentially influenced by the occurrence of recurrent events. The association among them is modeled by the Clayton copula. Thus, this thesis also introduces new features into the statistical literature through the YP_{EX} , YP_{PE} , YP_{BP} , PH_{PE} , PH_{BP} , PO_{EX} , PO_{PE} , and PO_{BP} joint frailty-copula models.

The models of the two classes include frailty terms, i.e., latent variables. In the frequentist approach, the likelihood function would need to be integrated with respect to this variable. For this reason, we choose to apply the Bayesian approach.

This thesis discusses some essential concepts of survival analysis, encompassing the PH, PO, and YP regressions, along with an overview of the frailty model. It further explored the Bernstein polynomials and the piecewise exponential, which were utilized for modeling baseline hazard functions. Additionally, the thesis examined the concepts and properties of copulas, going deeper specifically for the Clayton copula. We presented details of our proposed models, starting with the definition of notation and proceeding to elaborate on the construction of the likelihood function. The methodology for data generation and the findings from the Monte Carlo study were thoroughly presented and analyzed. Finally, the thesis presented the application of the models.

As for numerical results, we executed a Monte Carlo simulation study for each model class aimed at evaluating the influence of model selection on parameter estimation. This assessment focused on some criteria such as estimation biases (RB), average standard error (ASE), standard deviation of estimates (SDE), credible intervals, and coverage probability (CP). A total of $M_C = 250$ Monte Carlo replicas were generated, each comprising $L = 300$ individuals in 12 scenarios. In all the scenarios, our estimates mean and median are generally close to the true values, indicating a good level of accuracy. In addition, the ASE and SDE values are close and the CP values are not very far from 95%. These facts signal a good performance of our models.

In the data analysis, we fitted our models on the `readmission` database. This dataset contains the times to the death of patients with colorectal cancer along with the times between their hospital readmissions due to cancer. Three characteristics related to individuals were evaluated: sex, chemotherapy treatment, and Dukes' stage of cancer. All variables are categorical and showed no changes over the follow-up period. We initially only evaluate the time to terminal event by applying our first class of models. Then, we modeled the times to the terminal event jointly with the times between recurrent events by applying the second class of models. The performance of each adjustment was evaluated against the WAIC estimate. Concerning this criterion, the YP_{BP} models presented better values, in both the first and second class of models. We provide Shiny applications developed from the `Shiny` package (Chang et al., 2023) with all simulation

and real application results.

Some limitations of this research were:

- In our simulation studies, we did not apply WAIC to Monte Carlo samples. We acknowledge that the assessment of these values could enhance the depth of comparative analysis of our models.
- We did not consider Weibull baseline distributions in the fit of our models.
- In the real application, we believe that the high volume of administrative censoring somewhat reduced the predictive ability of our models.
- In the joint frailty-copula models, we only used the Clayton copula because it is the simplest and most used among the Archimedean copulas.
- Our models are not yet capable of accommodating time-dependent variables.

In future research, we want to apply the following approaches:

- Evaluate the WAIC of our model fits in a simulation study.
- Conduct a more extensive simulation study incorporating the baseline Weibull distribution.
- Incorporate a cure fraction model into our model classes.
- Incorporate other Archimedean copulas (Frank, Gumbel, Joe, and AMH) into the second class of models. It will allow us to deal with different correlation ranges, both positive and negative, and evaluate the impact of choosing the incorrect model on RB, ASE, SDE, CP, and WAIC.
- Adapt our regression frameworks to allow us to model time-dependent covariates.
- Extend our models to a frequentist approach.
- Deploy a residual analysis that provides an additional way of evaluating the quality of our fits.
- Publish an R package that provides the functions used in this thesis, facilitating the replication of its results.

References

- Akaike, H. (2011). *Akaike's Information Criterion*. In: Lovric, M. International Encyclopedia of Statistical Science. pages 25–25. Springer, Berlin, Heidelberg.
- Amorim, L. D. and Cai, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1):324–333.
- Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2):273–277.
- Bernstein, S. (1912). On the best approximation of continuous functions by polynomials of a given degree. *Kharkov Mathematical Society*, 2(13):49–194.
- Biondo, T. R. and Suzuki, A. K. (2016). Modelos de sobrevivência bivariados derivados da cópula arquimediana de clayton: Uma abordagem Bayesiana. *Matemática e Estatística em Foco*, 4(2):87–102.
- Breslow, N. (1972). Regression models and life-tables (by [Cox \(1972\)](#)). *Journal of the Royal Statistical Society: Series B*, 34(2):216–217.
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99.
- Chang, I.-S., Hsiung, C. A., Wu, Y.-j., and Yang, C.-c. (2005). Bayesian survival analysis using bernstein polynomials. *Scandinavian journal of statistics*, 32(3):447–466.
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2023). *shiny: Web Application Framework for R*. R package version 1.7.4.1.
- Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, 40(5):373–383.
- Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A*, 148(2):82–108.
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

- Collett, D. (2015). *Modelling survival data in medical research*. CRC Press, Boca Raton, Flórida, USA.
- Colosimo, E. A. and Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. Editora Blucher, São Paulo, SP, Brazil.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202.
- de Mello, J. F. (2016). Modelo exponencial por partes para dados de sobrevivência com longa duração. Master’s thesis, Universidade Federal de Minas Gerais, Statistics Department. Available online: <http://hdl.handle.net/1843/BUBD-AA2EHB> (accessed on 14 December 2023).
- Demarqui, F. N. (2020a). *Yang and Prentice Model with Baseline Distribution Modeled by Bernstein Polynomials*. R package version 0.0.1.
- Demarqui, F. N. (2020b). *YPPE: Yang and Prentice Model with Piecewise Exponential Baseline Distribution*. R package version 1.0.1.
- Demarqui, F. N., Dey, D. K., Loschi, R. H., and Colosimo, E. A. (2011). Modeling survival data using the piecewise exponential model with random time grid. *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, 4(1):109–122.
- Demarqui, F. N., Loschi, R. H., Dey, D. K., and Colosimo, E. A. (2012). A class of dynamic piecewise exponential models with random time grid. *Journal of Statistical Planning and Inference*, 142(3):728–742.
- Demarqui, F. N. and Mayrink, V. D. (2021). Yang and Prentice model with piecewise exponential baseline distribution for modeling lifetime data with crossing survival curves. *Brazilian Journal of Probability and Statistics*, 35(1):172–186.
- Demarqui, F. N., Mayrink, V. D., and Ghosh, S. K. (2019). An unified semiparametric approach to model lifetime data with crossing survival curves. *arXiv preprint arXiv:1910.04475*.
- Diao, G., Zeng, D., and Yang, S. (2013). Efficient semiparametric estimation of short-term and long-term hazard ratios with right-censored data. *Biometrics*, 69(4):840–849.
- Economou, P. and Caroni, C. (2007). Parametric proportional odds frailty models. *Communications in Statistics—Simulation and Computation*, 36(6):1295–1307.
- Emura, T., Nakatochi, M., Murotani, K., and Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research*, 26(6):2649–2666.

- Farouki, R. T. (2012). The bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419.
- Farouki, R. T. and Rajan, V. (1987). On the numerical condition of polynomials in bernstein form. *Computer Aided Geometric Design*, 4(3):191–216.
- Feller, W. (1987). *An introduction to probability theory and its applications*. John Wiley & Sons, Hoboken, Nova Jersey, USA.
- Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2):389–402.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC, New York, USA, first edition.
- González, J. R., Fernandez, E., Moreno, V., Ribes, J., Peris, M., Navarro, M., Cambray, M., and Borràs, J. M. (2005). Sex differences in hospital readmission among colorectal cancer patients. *Journal of Epidemiology and Community Health*, 59(6):506.
- Gupta, R. C. and Peng, C. (2014). Proportional odds frailty model and stochastic comparisons. *Annals of the Institute of Statistical Mathematics*, 66(5):897–912.
- Hanagal, D. D. (2011). *Modeling survival data using frailty models*. Springer, New York, USA, first edition.
- Hanson, T. and Yang, M. (2007). Bayesian semiparametric proportional odds models. *Biometrics*, 63(1):88–95.
- Hofert, M., Kojadinovic, I., Mächler, M., and Yan, J. (2018). *Elements of copula modeling with R*. Springer, New York, USA.
- Hosmer, D. W. and Lemeshow, S. (1999). *Applied survival analysis: Time-to-event*, volume 317. Wiley-Interscience, New York, USA.
- Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer, New York, USA.
- Huang, C.-Y. and Wang, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*, 99(468):1153–1165.
- Huang, X. and Liu, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics*, 63(2):389–397.
- Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics*, 58(3):510–520.

- Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2014). *Bayesian Survival Analysis*. Wiley, New York, USA.
- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.
- Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on cox’s regression and life model. *Biometrika*, 60(2):267–278.
- Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, Hoboken, Nova Jersey, USA.
- Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer, New York, USA.
- Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*. Springer, New York, USA.
- Koralov, L. and Sinai, Y. G. (2007). *Theory of probability and random processes*. Springer, Berlin, Germany.
- Lawless, J. F. (1987). Regression methods for poisson process data. *Journal of the American Statistical Association*, 82(399):808–815.
- Lawless, J. F. (2011). *Statistical models and methods for lifetime data*. John Wiley & Sons, Hoboken, Nova Jersey, USA.
- Li, Z., Chinchilli, V. M., and Wang, M. (2019). A bayesian joint model of recurrent events and a terminal event. *Biometrical Journal*, 61(1):187–202.
- Librero, J., Peiró, S., and Ordinaña, R. (1999). Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days. *Journal of Clinical Epidemiology*, 52(3):171–179.
- Lin, X. and Wang, L. (2011). Bayesian proportional odds models for analyzing current status data: univariate, clustered, and multivariate. *Communications in Statistics-Simulation and Computation*, 40(8):1171–1181.
- Liu, L., Wolfe, R. A., and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3):747–756.
- Lorentz, G. G. (1986). *Bernstein polynomials*. American Mathematical Soc., New York, USA.
- Mazroui, Y., Mathoulin-Pelissier, S., Soubeyran, P., and Rondeau, V. (2012). General joint frailty model for recurrent event data with a dependent terminal event: application to follicular lymphoma data. *Statistics in Medicine*, 31(11-12):1162–1176.

- Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science & Business Media, New York, USA.
- Ninomiya, Y. (2021). Prior intensified information criterion. *arXiv preprint arXiv:2110.12145*.
- Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B*, 44(3):414–422.
- Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493.
- Osman, M. and Ghosh, S. K. (2012). Nonparametric regression models for right-censored data using bernstein polynomials. *Computational Statistics and Data Analysis*, 56(3):559–573.
- Panaro, R., Demarqui, F., and Mayrink, V. (2020). *spsurv: Bernstein Polynomial Based Semiparametric Survival Analysis*. R package version 1.0.0.
- Panaro, R. V. (2020). *spsurv: An r package for semi-parametric survival analysis*. *arXiv preprint arXiv:2003.10548*.
- Patiño, E. G. (2018). *Modelo Bayesiano para dados de sobrevivência com riscos semi-competitivos baseado em cópulas*. PhD thesis, Universidade de São Paulo, Statistics Department. Available online: <https://teses.usp.br/teses/disponiveis/45/45133/tde-17072018-155825/pt-br.php> (accessed on 03 February 2022).
- Peng, R. D. (2020). *Advanced statistical computing*. Available online: <https://bookdown.org/rdpeng/advstatcomp/metropolis-hastings.html> (accessed on 14 October 2022).
- Prenen, L., Braekers, R., and Duchateau, L. (2017). Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society: Series B*, 79(2):483–505.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rondeau, V., Marzroui, Y., and Gonzalez, J. R. (2012). frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47:1–28.
- Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4):708–721.

- Royston, P. and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modeling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197.
- Schneider, S. (2017). *Modelos para dados de sobrevivência multivariados com censura informativa*. PhD thesis, Universidade Federal de Minas Gerais, Statistics Department. Available online: <http://hdl.handle.net/1843/BUBD-AXGPA3> (accessed on 02 June 2022).
- Schneider, S., Demarqui, F. N., Colosimo, E. A., and Mayrink, V. D. (2020). An approach to model clustered survival data with dependent censoring. *Biometrical Journal*, 62(1):157–174.
- Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.1.
- Stan Development Team (2023). Stan modeling language users guide and reference manual: Prior choice recommendations. Available online: <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations> (accessed on 10 May 2023).
- Suzuki, A. K. (2012). *Modelos de sobrevivência bivariados baseados na cópula FGM: uma abordagem bayesiana*. PhD thesis, Universidade Federal de São Carlos, Statistics Department. Available online: <https://repositorio.ufscar.br/handle/ufscar/4487> (accessed on 12 November 2022).
- Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2023). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.6.0.
- Vehtari, A., Gelman, A., Gabry, J., and Yao, Y. (2021). Package ‘loo’. *Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. R package version 2.6.0.
- Wang, L. and Dunson, D. B. (2011). Semiparametric bayes’ proportional odds models for current status data with underreporting. *Biometrics*, 67(3):1111–1118.
- Wang, Y. (2013). *Sample Size Calculation Based on the Semiparametric Analysis of Short-term and Long-term Hazard Ratios*. PhD thesis, Columbia University, Statistics Department. Available online: <https://academiccommons.columbia.edu/doi/10.7916/D8ST7X25> (accessed on 13 November 2022).
- Wienke, A. (2020). *Frailty models in survival analysis*. CRC Press, Boca Raton, Flórida, USA.

-
- Wong, J.-M., Yen, M.-F., Lai, M.-S., Duffy, S. W., Smith, R. A., and Chen, T. H.-H. (2004). Progression rates of colorectal cancer by dukes' stage in a high-risk group: analysis of selective colorectal cancer screening. *The Cancer Journal*, 10(3):160–169.
- Yang, S. and Prentice, R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*, 92(1):1–17.
- Yang, S. and Zhao, Y. (2012). Checking the short-term and long-term hazard ratio model for survival data. *Scandinavian Journal of Statistics*, 39(3):554–567.

Appendix A

Numerical and graphical results of all models

A.1 Monte Carlo simulation study

The results of the Monte Carlo simulation study of all models are available in the Shiny application. These applications can be accessed via the links

- cassiushenrique.shinyapps.io/appSimulationsFrailty, and

Figure A.1: Numerical and graphical results of the Monte Carlo simulation study of the models of the first class of models.



Source: Prepared by the author.

- cassiushenrique.shinyapps.io/appSimulationsJointFrailtyCopula

Figure A.2: Numerical and graphical results of the Monte Carlo simulation study of the models of the second class of models.



Source: Prepared by the author.

or by the QR codes shown in the Figures [A.1](#) (YP frailty models) and [A.2](#) (the joint frailty-copula models).

A.2 Real application

The outcomes from the real application of all models can also be explored through a Shiny application. Access to these applications is provided via the links

- cassiushenrique.shinyapps.io/appRealFrailty, and

Figure A.3: Numerical and graphical results of the real application of the models of the first class of models.



Source: Prepared by the author.

- cassiushenrique.shinyapps.io/appRealJointFrailtyCopula

Figure A.4: Numerical and graphical results of the real application of the models of the second class of models.



Source: Prepared by the author.

or by the QR codes shown in the Figures [A.3](#) (YP frailty models) and [A.4](#) (the joint frailty-copula models).