KGsurv: an R package to fit the Kumaraswamy-G family of distributions for survival data

Caio Gabriel Barreto Balieiro

Departamento de Estatística - ICEx - UFMG

Maio de 2021

Caio Gabriel Barreto Balieiro

KGsurv: an R package to fit the Kumaraswamy-G family of distributions for survival data

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Fábio Nogueira Demarqui.

Belo Horizonte, MG - Brasil
Maio de 2021

Balieiro, Caio Gabriel Barreto

B186k KGsurv [manuscrito]: an R package to fit the Kumaraswamy-G family of distributions for survival data / Caio Gabriel Barreto Balieiro. – 2021.
xi, 84 f. il.

Orientador: Fábio Nogueira Demarqui.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 80–84.

1. Estatística – Teses. 2. Análise de sobrevivência (Biometria) – Teses. 3. Análise de regressão – Teses. I. Demarqui, Fábio Nogueira. II Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV.Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ATA DA DEFESA DE DISSERTAÇÃO DE MESTRADO DO ALUNO CAIO GABRIEL BARRETO BALIEIRO, MATRICULADO, SOB O Nº 2019661378, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 13 DE MAIO DE 2021.

Aos 13 dias do mês de Maio de 2021, às 14h00, em reunião pública virtual 259 (conforme orientações para a atividade de defesa de dissertação durante a vigência da Portaria PRPG nº 1819) no Instituto de Ciências Exatas da UFMG, https://meet.google.com/gpu-cdgm-jjm, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de dissertação do(a) aluno CAIO GABRIEL BARRETO BALIEIRO, nº matrícula 2019661378, intitulada: "*KGsurv: an R package to fit the Kumaraswamy-G family of distributions for survival data*", requisito final para obtenção do Grau de mestre em Estatística. Abrindo a sessão, o(a) Senhor(a) Presidente da Comissão, Prof. Fábio Nogueira Demarqui, passou a palavra ao(à) aluno(a) para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovada.
( ) Reprovada com resubmissão do texto em ____ dias.
( ) Reprovada com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovada.

______________________	Cristiano de Carvalho Santos
Prof. Fabio Nogueira Demarqui – Orientador (EST/UFMG)	Prof. Cristiano de Carvalho Santos (EST/UFMG)
______________________	______________________
Prof. Enrico Antonio Colosimo (EST/UFMG)	Prof. Jeremias da Silva Leão (EST/UFAM)
______________________	______________________

O resultado final foi comunicado publicamente ao(à) aluno(a) pelo(a) Senhor(a) Presidente da Comissão. Nada mais havendo a tratar, o(a) Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 13 de maio de 2021.

Observações:
1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá

Aos meus avôs Cosmo Balieiro e Luiz Leal (em memorial).

“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”.

George E. P. Box

Agradecimentos

Primeiramente quero agradecer a Deus que permitiu-me estudar na UFMG. Através dEle, conheci pessoas importantes que ajudaram-me a melhorar meu conhecimento à cerca da Estatística.

Gostaria de agradecer a minha família por todo o suporte, principalmente meus pais Elcirene Barreto, Antônio Balieiro e Dimas Leal (padrasto) que acreditaram em meu potencial e me doaram todo o suporte financeiro para que pudesse viajar para outro estado. Além disso, quero agradecer meu irmão Luiz Borges por todo o suporte de manutenção dos meus computadores e todas as dicas computacionais para que pudesse ter o maior desempenho nas minhas análises.

À minha namorada Gabriela Guedes e sua família que me ajudaram bastante no processo de adaptação em uma cidade completamente nova para mim. Deixo meus agradecimentos a Adriana Guedes e Araci Guedes. Quero também agradecer aos amigos que fiz em Belo Horizonte, sendo eles: Angela, Alexander, César, e Walmir.

Um dos pontos positivos do meu mestrado foi encontrar um orientador que confiou em mim a partir da nossa primeira conversa. Deixo aqui meus sinceros agradecimentos ao meu orientador Fábio Nogueira Demarqui que fez-me evoluir bastante e principalmente me deu todo o suporte necessário para que eu pudesse apresentar o meu máximo neste documento.

Gostaria de agradecer aos professores membros da banca Cristiano de Carvalho Santos, Enrico Antônio Colosimo, e Jeremias da Silva Leão pelas imensas contribuições para a melhora deste exemplar.

Por fim, quero agradecer a FAPEMIG e a CAPES o suporte financeiro através de uma bolsa de estudos, o que tornou possível meu sonho de estudar na UFMG.

Resumo

Neste trabalho, foi proposto um novo pacote em linguaguem R (denominado KGsurv) utilizando algoritmos implementados no software Stan, para modelar novos modelos baseados na família de distribuições Kumaraswamy-G. Nós apresentamos os modelos de regressão Kumaraswamy-G para as seguintes classes: riscos proporcionais, chances proporcionais e tempo de vida acelerado considerando as distribuições Exponencial, Weibull, Gamma, Log-logística, e Log-normal para modelar a distribuição de G. Neste trabalho, a abordagem frequentista foi considerada para estimar os parâmetros dos modelos com dados de sobrevivência censurados à direita sob o pressuposto de um mecanismo de censura não informativo. Por fim, foram apresentadas aplicações utilizando três conjuntos de dados reais já utilizados na literatura de análise de sobrevivência para verificar os resultados dos modelos implementados no pacote KGsurv.

Palavras-chave: Kumaraswamy-G, análise de sobrevivência, modelos de regressão, pacote KGsurv.

Abstract

In this work, a new package in language R (called KGsurv) was proposed using algorithms implemented in Stan software, to model new models based on the Kumaraswamy-G family of distributions. We present the Kumaraswamy-G regression models for the following classes: proportional hazards, proportional odds, and accelerated failure time considering Exponential, Weibull, Gamma, Log-logistics, and Log-normal distributions to model the G distribution. In this work, the frequentist approach was considered to estimate the parameters of the models with right-censored survival data under the assumption of a non-informative censoring mechanism. Finally, applications were presented using three real data sets widely used in the survival analysis literature to verify the results of the models implemented in the KGsurv package.

Keywords: Kumaraswamy-G, survival analysis, regression models, KGsurv package.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The analysis of survival and reliability data has gained much space in the literature due to a range of new models that can be used in different areas of knowledge. Thus, this is an important tool for statistical analysis. In survival and reliability analysis we are generally interested in the time until the occurrence of the event of interest, such as the time to death of patients diagnosed with a certain type of cancer, the time until the failure of a lamp, or the time until customers close their bank accounts, among others.

The main characteristic of survival data is the presence of incomplete or partial observations. These observations, known as censoring, can occur for several reasons, such as the patient's death for a different cause from the one studied, loss of follow-up before the end of the study, the patient's moves to another city, among others (Colosimo and Giolo, 2006).

The usual theory of survival analysis assumes that, if observed for a long period, all individuals will fail at some point, that is, they are susceptible to the event of interest. Also, in many studies, there are explanatory variables that are related to the time of failure. Therefore, we must use regression models in the context of survival analysis. The most famous regression model is the proportional hazards (PH) model defined by Cox (1972). This model allows the incorporation of covariates through the hazard function. However, it has the assumption of proportional hazards. Thus, in many situations, this model cannot be used. The second alternative is the proportional odds (PO) model defined in Bennett (1983). This model is not used frequently in the literature. However, this model presents an attractive interpretation of the parameters, using the odds ratio. Besides, the PO model has the assumption that the odds functions must be proportional. Finally, the accelerated failure time model (AFT) is widely used in the medical and engineering literature. Furthermore, the AFT model does not have the assumption of proportionality of hazard or odds. Then, this model is a good alternative. For more details, see Lawless (2011).

For instance, in survival analysis, the main package is the so-called `survival` proposed by (Therneau, 2014), this package uses as main techniques of survival analysis, for example, Kaplan-Meier survival curves, Logrank test, and PH and AFT models. A second package that can be used in survival analysis is `flexsurv` introduced in (Jackson, 2016), this package uses the AFT models considering several baseline distributions, such as Exponential, Weibull, Log-normal, Log-logistics, among others. To use the PO model on the R platform, you can use the `timereg` package (Scheike and Zhang, 2011). However, the results obtained in the `timereg` package are quite limited, for example, this package does not have functions for viewing the survival curves determined by the PO model, that is, the verification of the adequacy of the model fit cannot be verified. It is important to highlight that, there are other packages that use the PH, PO, or AFT models, for instance, `rms` (Harrell Jr et al., 2016), `SurvRegCensCov` (Hubeaux and Rufibach, 2014), and `eha` (Brostom, 2014) considers the accelerated failure time models, but implemented differently the `survival` package. Also, the `spsurv` package presented by Panaro (2020), uses the PH, PO, and AFT models with the Bernstein polynomials as a baseline distribution, among others.

In the present work, a new package R was presented, using the software Stan Team (2018), this package was called `KGsurv`. In R, the Stan software is implemented in `rstan` package. The `KGsurv` package was created with the objective is to use the families of regression models PH, PO, and AFT using easy routines to obtain the estimation of parameters of these models. The `KGsurv` package uses the frequentist approach considering the censoring mechanism is non-informative, and right censoring. To model the baseline distribution, the Kumaraswamy-G family of distributions (Kumaraswamy-G) presented in Cordeiro et al. (2010) and Cordeiro and de Castro (2011) was considered. This distribution is very flexible, that is, it has several forms of density, survival, and hazard functions. Besides that, several distributions have been created based on the Kumaraswamy-G family of distributions, such as the Kumaraswamy Generalized Gamma (De Pascoa et al., 2011), the Kumaraswamy Birnbaum-Saunders (Saulo et al., 2012), the Kumaraswamy generalized half-normal distribution (Cordeiro et al., 2012), the Kumaraswamy Burr XII distribution (Paranaíba et al., 2013), the Kumaraswamy exponentiated Pareto distribution (Elbatal, 2013), the Kumaraswamy half-Cauchy distribution (Ghosh, 2014), the Kumaraswamy modified Weibull (Cordeiro et al., 2014), among others.

The Kumaraswamy-G family of distributions is widely used in the literature, so presenting a package in the language R which uses this distribution with a friendly routine and easy access makes the `KGsurv` package an attractive alternative to survival analysis data with the right-censoring.

1.1 Objectives of the dissertation

The general objective of this work is to build a package in language R to fit the regressions families PH, PO, and AFT considering the Kumaraswamy-G family of distributions as the baseline distribution. Our specific objectives are:

- To introduce the Kumaraswamy-G family of distributions with the PH, PO, and AFT models, considering the Exponential, Weibull, Gamma, Log-logistic, and Log-normal distributions to model G distribution.

- To present an R package, called `KGsurv`, to model the Kumaraswamy-G family of distributions for survival data with right-censored.

- To present three studies using real data sets to show that the new `KGsurv` package provides similar results with models already presented in the literature.

1.2 Organization of the chapters

This present work is organized as follows. In Chapter 2, we present the basic concepts of survival analysis. In addition, we present the families of regression models PH, PO, and AFT. In Chapter 3, we presented the Kumaraswamy-G family of distributions in the context of the survival analysis and we define the Kumaraswamy-G family of distributions considering the Exponential, Weibull, Gamma, Log-logistic, and Log-normal distribution to model the G distribution. Next, we introduced the families of regression models PH, PO, and AFT whose baseline function is the Kumaraswamy-G family of distributions. Lastly, we present the inferential procedures, the model selection criteria, and one little discussion of some functions in the `KGsurv` package. In Chapter 4, we showed three applications using real data sets. Finally, in Chapter 5, we presented the conclusions of the work and future applications.

Chapter 2

Survival analysis

In this chapter, we revisit some basic concepts in survival analysis. After that, we define the proportional hazard, proportional odds, and accelerated failure time regression models.

2.1 Basic concepts

In survival analysis, the response variable is the time until the occurrence of an event of interest usually referred to as failure time or lifetime. In clinical trials, some examples of failure times include time to death, cure, or recurrence of disease in patients. In engineering, studies are very common in which products, components, or systems are tested to study their reliability (this area is known as Reliability). For the interested reader, see Nelson (1990), Meeker and Escobar (2014). In financial data, we can study the time until customers leave a bank Hoggart and Griffin (2001). In this way, survival analysis can be applied in many areas of knowledge.

Survival data requires special treatment, because survival times are non-negative, and usually are governed by skewed distributions. Another peculiarity of survival data regards the presence of incomplete observations, known as censored times. Censored observations may occur due to several causes, for example, limitation of time or resources available for the study and loss of follow-up of a patient before the end of study (Colosimo and Giolo, 2006).

Survival data is said to be left-censored when the event of interest is known to have occurred before a certain time t, but the exact time of the occurrence of the event of interest is unknown. Interval-censored survival data arises when the event of interest is only known to have occurred in a given time interval [u, l]. Finally, we say that survival data is right-censored when there is a loss of follow-up or non-occurrence of the event of interest during the observation period, or before the study ends. We can divide it into three different types:

- Type I: occurs when the study is designed to end after a certain follow-up time. In this case, the number of failures is random.
- Type II: occurs when the study is terminated after a certain preestablish number of failures is reached. In this case, the follow-up time is random.
- Type III (random): occurs when an individual is withdrawn from the study without the failure, or also, for example, if the individual dies for a different reason than the one studied Colosimo and Giolo (2006).

For the interested reader, other practical examples along with a deeper discussion regarding censoring in the context of survival are presented in Lawless (2011). The right censoring scheme occurs more frequently in practice, for this reason, the models described in this work are for right-censored data.

Let T and C be random variables representing the time to failure and censoring, respectively. The right-censored survival data are characterized by

$$Y_i = \min\{T_i, C_i\} \quad \text{and} \quad \delta_i = \begin{cases} 1, \text{If } \ T_i \leq C_i \\ 0, \text{If } \ T_i > C_i, \end{cases}$$

where to each individual, $i = 1, 2, \ldots, n$, we have the pair (Y_i, δ_i), Y_i is the observable time and δ_i is the censoring indicator. It is worth mentioning that when T and C are independent, we say that the censoring mechanism is non-informative, otherwise, we say that the censoring mechanism is informative.

Let T be a non-negative random variable denoting the time until an occurrence of an event of interest. Define $\boldsymbol{\zeta}$ as a vector of parameters, so the cumulative distribution function (c.d.f.) is expressed by

$$P(T \leq t) = F(t|\boldsymbol{\zeta}) = \int_0^t f(u|\boldsymbol{\zeta})\text{du}, \ \ \text{t} > 0,$$

where $f(\cdot|\boldsymbol{\zeta})$ is a probability density function (p.d.f.).

The survival function is defined to be the probability that the survival time is greater than to t, and so

$$S(t|\boldsymbol{\zeta}) = P(T > t) = 1 - F(t|\boldsymbol{\zeta}), t > 0.$$

The survival function has some important properties:

(i) $S(t|\boldsymbol{\zeta})$ it is an non-increasing function of t;

(ii) $S(0) = 1$;

(iii) $\lim_{t \to \infty} S(t|\boldsymbol{\zeta}) = 0$.

According to Colosimo and Giolo (2006) the failure (or hazard) rate function is more informative than the survival function because different survival functions can have similar shapes, while the failure rate functions can differ dramatically. This function is expressed by

$$h(t|\boldsymbol{\zeta}) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t|\boldsymbol{\zeta})}{S(t|\boldsymbol{\zeta})}.$$

The cumulative hazard function is defined as

$$H(t|\boldsymbol{\zeta}) = \int_0^t h(u|\boldsymbol{\zeta}) du. \tag{2.1}$$

The odds function can be expressed as

$$R(t|\boldsymbol{\zeta}) = \frac{F(t|\boldsymbol{\zeta})}{S(t|\boldsymbol{\zeta})}. \tag{2.2}$$

The functions described above are mathematically related, therefore satisfy the following relationships:

$$f(t|\boldsymbol{\zeta}) = -\frac{d}{dt} S(t|\boldsymbol{\zeta}) = -S'(t|\boldsymbol{\zeta}),$$

$$h(t|\boldsymbol{\zeta}) = \frac{-S'(t|\boldsymbol{\zeta})}{S(t|\boldsymbol{\zeta})} = -\frac{d}{dt}\left[\log(S(t|\boldsymbol{\zeta}))\right], \tag{2.3}$$

$$H(t|\boldsymbol{\zeta}) = \int_0^t h(u|\boldsymbol{\zeta})du = -\log\left[S(t|\boldsymbol{\zeta})\right]. \tag{2.4}$$

$$S(t|\boldsymbol{\zeta}) = \exp\left[-H(t|\boldsymbol{\zeta})\right] = \exp\left[-\int_0^t h(u|\boldsymbol{\zeta})du\right]. \tag{2.5}$$

$$R(t|\boldsymbol{\zeta}) = \exp\left[H(t|\boldsymbol{\zeta})\right] - 1. \tag{2.6}$$

To obtain inference about the vector of parameters $\boldsymbol{\zeta}$, we will assume that for right-censored survival data, and under the assumption of a non-informative censoring mechanism, we can write the likelihood function as (Lawless, 2011, p. 55)

$$\begin{aligned} L(\boldsymbol{\zeta}|\mathcal{D}) &= \prod_{i=1}^{n} \left[f(y_i|\boldsymbol{\zeta})\right]^{\delta_i} \left[S(y_i|\boldsymbol{\zeta})\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[h(y_i|\boldsymbol{\zeta})\right]^{\delta_i} S(y_i|\boldsymbol{\zeta}), \end{aligned} \tag{2.7}$$

where $\mathcal{D} = (y_i, \delta_i, i = 1, 2, \ldots, n)$ is denotes the observed data.

From the frequentist point of view, to find the maximum likelihood (ML) estimators to the model, we define $l(\boldsymbol{\zeta}|\mathcal{D}) = \log(L(\boldsymbol{\zeta}|\mathcal{D}))$, where $L(\boldsymbol{\zeta}|\mathcal{D})$ is given in (2.7), after that, maximizes the function, that is, finds the resolution estimators or system of equations is given by

$$U(\boldsymbol{\zeta}|\mathcal{D}) = \frac{\partial l(\boldsymbol{\zeta}|\mathcal{D})}{\partial \boldsymbol{\zeta}} = \mathbf{0},$$

where $U(\cdot)$ is called the Score function. In problems with high parameter dimensions, these forms are closed to the equation (2.8) are not viable, that is, we should use numerical methods to estimate the parameters, for example, the Newton-Rapson algorithm.

In many experiments, it is interesting to evaluate the interval estimates of the parameters. Consider a $\boldsymbol{\zeta}$ vector of parameters of interest of size q. Thus, using the ML properties and under standard regularity conditions, see Cox and Hinkley (1979), we have

$$\widehat{\boldsymbol{\zeta}} \approx N_k(\boldsymbol{\zeta}, \mathcal{I}^{-1}(\boldsymbol{\zeta})),$$

where $k = 1, 2, \dots, q$, and $\mathcal{I}(\boldsymbol{\zeta}) = -E\left[\frac{\partial^2}{\partial\boldsymbol{\zeta}\boldsymbol{\zeta}^\top} l(\boldsymbol{\zeta}|\mathcal{D})\right]$. However, calculating this Fisher information in practice in many problems is impracticable, as it depends on the parameter of interest. Therefore, we can use the observed Fisher information, which is given by the following Equation

$$\mathcal{F}(\widehat{\boldsymbol{\zeta}}) = \left[\frac{\partial^2}{\partial\boldsymbol{\zeta}\boldsymbol{\zeta}^\top} l(\boldsymbol{\zeta}|\mathcal{D})|_{\zeta=\hat{\zeta}}\right].$$

For the construction of asymptotic confidence intervals for the components of the $\boldsymbol{\zeta}$ vector, we can calculate in a usual way, that is

$$IC[\boldsymbol{\zeta}_k; \times 100(1-\alpha)\%] = \hat{\zeta}_k \pm z_{\alpha/2}\sqrt{\mathcal{F}_{kk}(\widehat{\boldsymbol{\zeta}})},$$

where α is the level of significance.

Besides that, to the point and interval estimation, we considered the hypothesis tests, namely: the Wald test (Wald, 1943). Usually, to test the hypothesis that the parameters are significant. Under $H_0 : \zeta = \zeta_0$ and the test statistic used is given by

$$W = (\hat{\zeta} - \zeta_0)^\top(-\mathcal{F}(\zeta_0))(\hat{\zeta} - \zeta_0).$$

Under H_0 the statistic test has an approximate chi-square distribution with q degrees of freedom.

In the next subsection, we present a brief overview of the proportional hazards, proportional odds, and Accelerated failure time models, discussing its main functions and interpretation of the parameters.

2.2 Regression models

In many studies, there are covariates or explanatory variables that can be related to the lifetime of patients, equipment, individuals, etc. For instance, the time until the death of cancer patients, some possible covariates are sex, age, etc. In engineering, the time until the failure of the equipment, the covariates can be the year of manufacture, the type of the material used, among others. In this section, we present three families of regression models in survival analysis, namely PH, PO, and AFT regression models.

2.2.1 Proportional hazards model

One of the most famous regression models in survival analysis is the model proposed by Cox (1972), called proportional hazards models. This model relates the hazard function in a multiplicative way to the effect of the explanatory variables. The hazard function of the PH model can be expressed by

$$h(t|\boldsymbol{\beta}, \mathbf{x}) = h_0(t)e^{\mathbf{x}_i^\top \boldsymbol{\beta}}, \tag{2.8}$$

where $h_0(t)$ is called a baseline hazard function, because when $\mathbf{x} = \mathbf{0}$ we have $h(t|\boldsymbol{\beta}, \mathbf{x}) = h_0(t)$, $\mathbf{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a vector of exploratory variables or covariates and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a vector of the regression coefficients.

In Cox's original formulation the hazard function in Equation (2.8), is composed by a non parametric component, $h_0(t)$, and a parametric component, $e^{\mathbf{x}_i^\top \boldsymbol{\beta}}$, and the estimation of the regression coefficients $\boldsymbol{\beta}$ is carried out through the partial likelihood Colosimo and Giolo (2006). A fully parametric approach for the PH model can also be obtained by specifying parametric baseline hazard functions, such as the Exponential, Weibull, Gamma, Log-logistic, Log-normal distributions, among others.

From the Equation (2.5), the survival function associated with the PH model is given by

$$S(t|\boldsymbol{\beta}, \mathbf{x}) = S_0(t)^{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} = \exp\{-H_0(t)e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\}, \tag{2.9}$$

where $S_0(t)$ and $H_0(t)$ are the baseline survival and cumulative hazard functions, respectively.

The PH model allows an attractive interpretation in terms of its hazard ratio (HR), given two individuals i and j their hazards ratio is express by

$$\text{HR} = \frac{h_0(t)e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{h_0(t)e^{\mathbf{x}_j^\top \boldsymbol{\beta}}} = \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_j^\top \boldsymbol{\beta}\}. \tag{2.10}$$

Note that the Equation (2.10) does not depend on time. For this reason, the PH model is known in the literature as the proportional hazards model.

The proportional hazards assumption plays a central role in the goodness of fit of the PH model. According to Struthers and Kalbfleisch (1986), ignoring this assumption can lead to bias in the estimation of the model coefficients. To verify that the model does not violate this assumption (Colosimo and Giolo, 2006) suggest the use of scaled Schoenfeld residuals presented in Schoenfeld (1982). These residuals can be calculated in R language using the `survival::cox.zph` or `survminer::ggcoxzph` functions.

In this work, we are concerned with a fully parametric specification of the PH model. Then, for right-censored survival data and under the assumption of a non-informative censoring mechanism, the likelihood function can be written from the Equation (2.7) as follows

$$\begin{aligned} L(\boldsymbol{\zeta}, \boldsymbol{\beta}|D) &= \prod_{i=1}^{n} \left[h(y_i|\boldsymbol{\zeta}, \boldsymbol{\beta}, \mathbf{x})\right]^{\delta_i} S(y_i|\boldsymbol{\zeta}, \boldsymbol{\beta}, \mathbf{x}) \\ &= \prod_{i=1}^{n} \left[h_0(y_i|\boldsymbol{\zeta}) e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\right]^{\delta_i} \exp\{-H_0(y_i|\boldsymbol{\zeta}) e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\}, \end{aligned}$$

where $\boldsymbol{\zeta}$ is the vector of parameters associated with the baseline distribution, and $D = (y_i, \delta_i, \mathbf{x}_i, i = 1, 2, \ldots, n)$ denotes the observed data.

There are some alternatives available in the literature when the proportional hazards assumption is not valid, such as the PO and AFT models, which will be discussed in the next sections.

2.2.2 Proportional odds model

The PO model originally introduced by Bennett (1983) is a regression survival model. According to Bennett (1983), the PO model is structurally similar to the proportional hazards model of Cox and may be used in similar situations.

Although the PO model presents an attractive alternative to the PH model, according to Collett (2015), there are two reasons why this model was not widely used in practice. First, there are few routines available to fit the PO model, for instance, in language R, there are the `timereg`, and `spsurv` packages. The second is that the model is likely to give similar results to a Cox regression model that includes a time-dependent variable to produce non-proportional hazards (Collett, 2015).

The odds function this model can be expressed by

$$R(t|\boldsymbol{\beta}, \mathbf{x}) = R_0(t) e^{\mathbf{x}_i^\top \boldsymbol{\beta}}, \tag{2.11}$$

where $R_0(t)$ is the baseline odds functions. Given the Equations (2.3) e (2.11), we have that the hazard function is given by

$$h(t|\boldsymbol{\beta}, \mathbf{x}) = \frac{R_0'(t)e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + R_0(t)e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}, \tag{2.12}$$

where $R_0'(t) = \dfrac{dR_0(t)}{dt} = \dfrac{h_0(t)}{S_0(t)}$. From the Equation (2.2), we have that the survival function is given by

$$S(t|\boldsymbol{\beta}, \mathbf{x}) = \left[\frac{1}{1 + R_0(t)e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}\right]. \tag{2.13}$$

Similar to the PH model, the PO model allows an easy interpretation of the regression coefficients in terms of the odds ratio (OR). Specifically, the odds ratio for two individuals i and j is expressed by

$$\text{OR} = \frac{R_0(t)\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{R_0(t)\exp\{\mathbf{x}_j^\top \boldsymbol{\beta}\}} = \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_j^\top \boldsymbol{\beta}\}.$$

This model is known as the proportional odds PO model because the odds ratio associated with any elements is constant over time. Besides, the survival function is the probability that the survival time is greater than t. Therefore, according to Panaro (2020), if the OR is equal to 1, it indicates that the event understudy has the same probability of occurring in both groups. On the other hand, if an OR is greater than 1, it indicates that the event is less likely to occur in the reference group or baseline group. Finally, an OR less than 1 indicates that it is more likely to occur in the reference group.

The assumption of a constant odds ratio plays an important role in the goodness of fit of the PO model and should be checked in practice. Unfortunately, to the best of our knowledge, there are no tests implemented in R that can be used to check such assumptions.

In this work, the odds function $R_0(t)$ will be modeled parametrically. Therefore, from (2.7), the likelihood function for the PO model can be expressed as

$$\begin{aligned} L(\boldsymbol{\zeta}, \boldsymbol{\beta}|D) &= \prod_{i=1}^{n} [h(y_i|\boldsymbol{\zeta}, \boldsymbol{\beta}, \mathbf{x})]^{\delta_i} S(y_i|\boldsymbol{\zeta}, \boldsymbol{\beta}, \mathbf{x}) \\ &= \prod_{i=1}^{n} \left[\frac{R_0'(y_i|\boldsymbol{\zeta})e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + R_0(y_i|\boldsymbol{\zeta})e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}\right]^{\delta_i} \left[\frac{1}{1 + R_0(y_i|\boldsymbol{\zeta})e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}\right]. \end{aligned} \tag{2.14}$$

Two models were presented, namely PH and PO. As previously mentioned, the PH model is well known in survival analysis, however, it presents the proportional hazard function assumption, so one solution is to use the PO model, but this model is little used in the literature due to the few functions implemented in statistical software and also it presents the proportional odds function assumption. A solution for these models is called the AFT model. This model is widely used in the medical and engineering literature, due to its flexibility using several parametric distributions to model the baseline distributions. Next, the AFT model will be presented.

2.2.3 Accelerated failure time model

The PH model is the most popular regression model in survival analysis, but it can only be used in situations in which the proportional hazards assumption holds. An alternative to the PH model is the AFT model, which shall be discussed in this section.

The AFT model corresponds to a regression survival model, in which explanatory variables measured on an individual are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis (Collett, 2015). The general model of accelerated failure time is defined by

$$T_i = \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}\nu_i, \quad i = 1, 2, \ldots, n, \tag{2.15}$$

where T_i is the response variable and ν_i is stochastic component of the model, under the assumption that $\nu_1, \nu_2, \ldots, \nu_n$ are independent and identically distributed with the baseline survival function $S_0(\nu)$. It is important to highlight that, the model can be used on the original scale presented in Equation (2.15), or considering the logarithm of the function presented in Equation (2.15). Several distributions can be considered, for instance, if we consider the Weibull, Log-normal, Gamma, and Log-logistic distributions for T, we have the following distributions for ν Extreme value, Normal, Log-Gama, and Logistic, respectively.

The survival function for this model can be express by

$$\begin{aligned} S(t|\boldsymbol{\beta}, \mathbf{x}_i) &= P(T_i > t) = P\left(\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}\nu_i > t\right) \\ &= P\left(\nu_i > \frac{t}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right) \\ &= S_0\left(\frac{t}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right). \end{aligned} \tag{2.16}$$

The hazard function of this model, can be obtained from Equations (2.3) and (2.16) as follows

$$\begin{aligned} h(t|\boldsymbol{\beta}, \mathbf{x}_i) &= -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t|\mathbf{x}_i) \\ &= \frac{1}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} h_0\left(\frac{t}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right). \end{aligned}$$

The AFT model presents an easy interpretation of the regression coefficients in terms of a ratio involving the median survival times (or any other percentile). To obtain the time ratio (TR) between two individuals i and j it is necessary to obtain the survival percentiles if p is such that

$$p = S(t_p|\boldsymbol{\beta}, \mathbf{x}_i) = S_0\left(\frac{t_p}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right).$$

Then, we have

$$t_p(\boldsymbol{\beta}, \mathbf{x}_i) = S_0^{-1}(p) \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}.$$

Thus, any percentile ratio shall satisfy

$$\mathrm{TR} = \frac{t_p(\boldsymbol{\beta}, \mathbf{x}_i)}{t_p(\boldsymbol{\beta}, \mathbf{x}_j)} = \frac{S_0^{-1}(p) \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{S_0^{-1}(p) \exp\{\mathbf{x}_j^\top \boldsymbol{\beta}\}} = \exp\{\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_j^\top \boldsymbol{\beta}\}.$$

For the construction of the likelihood function of this model, a parametric model was be assumed to the model of baseline distribution. Then, for right-censored survival data, and under the assumption of non-informative censoring mechanism, the likelihood function can be rewritten from the Equation (2.7) as follows

$$\begin{aligned} L(\boldsymbol{\zeta}, \boldsymbol{\beta}|D) &= \prod_{i=1}^{n} \left[h(y_i|\boldsymbol{\zeta}, \boldsymbol{\beta}, \mathbf{x})\right]^{\delta_i} S(y_i|\boldsymbol{\zeta}, \boldsymbol{\beta}, \mathbf{x}) \\ &= \prod_{i=1}^{n} \left[\frac{1}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} h_0\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]^{\delta_i} S_0\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right). \end{aligned} \tag{2.17}$$

The three families of regression models presented above involve a baseline distribution, which can be modeled in a parametric or non-parametric way. Thus, many distributions can be used to model the baseline distribution. In this work, we used the Kumaraswamy-G family of distributions defined in Cordeiro et al. (2010) and Cordeiro and de Castro (2011), the results will be presented in the next Chapter.

Chapter 3

Kumaraswamy-G family of distributions in survival analysis

In this chapter, we present the Kumaraswamy-G family of distributions. Next, we present some particular cases for this family such as Kumaraswamy-Exponential, Kumaraswamy-Weibull, Kumaraswamy-Gamma, Kumaraswamy-Log-Logistic, Kumaraswamy-Log-normal distributions. Besides, we presented the proportional hazards regression models with Kumaraswamy-G baseline distribution (Kum-G-PH), the proportional odds regression models with Kumaraswamy-G baseline distribution (Kum-G-PO), and the accelerated failure time regression models with Kumaraswamy-G baseline distribution (Kum-G-AFT). Finally, we present the inferential procedures, the model selection criteria, and one little discussion of some functions in the package `KGsurv`.

3.1 Kumaraswamy-G family of distributions

Kumaraswamy (1980) introduced a two-parameter continuous distribution with support on (0, 1), which is the so-called Kumaraswamy distribution. The author presents some properties and applications of this model. Assunção (2018) used the Kumaraswamy distribution in quantile spatial regression to predict the wind speed.

The cumulative distribution and probability density functions associated with the

Kumaraswamy distribution are given by

$$F(t|a,b) = 1-(1-t^{a})^{b},\ \ t\in(0,1), \tag{3.1}$$

and

$$f(t|a,b) = abt^{a-1}(1-t^{a})^{b},\ \ t\in(0,1), \tag{3.2}$$

where $a, b > 0$. For the interested reader in the properties of the Kumaraswamy distribution, see Kumaraswamy (1980) and Jones (2009), for more details. Another distribution defined in the interval (0,1) is the Beta distribution, its the p.d.f. is express by

$$f(t|a,b) = \frac{1}{B(a,b)}t^{a-1}(1-t)^{b-1},\ \ t\in(0,1), \tag{3.3}$$

where $a, b > 0$ and $B(\cdot,\cdot)$ is the beta function. Note that, the Kumaraswamy density function is simpler than the Beta density function (it does not depend on the beta function). Besides that, according to Jones (2009), the Kumaraswamy distribution has many advantages over the beta distribution. For instance, the quantile function, the random variate generation, the moments, and the order statistics are available in simple forms.

The Kumaraswamy-G (hereafter Kum-G) family of distributions were presented in Cordeiro et al. (2010) and Cordeiro and de Castro (2011). According to Cordeiro and de Castro (2011) the idea of creating the Kum-G distribution came from the results presented in the class of generalized beta distributions Eugene et al. (2002) and Jones (2009). Consider an arbitrary baseline c.d.f. $G(t|\boldsymbol{\zeta})$, so $g(t|\boldsymbol{\zeta}) = \dfrac{dG(t|\boldsymbol{\zeta})}{dt}$ is the (p.d.f.). Define $\boldsymbol{\gamma} = (a, b, \boldsymbol{\zeta})$, where $\boldsymbol{\zeta}$ is the vector of baseline parameters. Then, the cumulative distribution and probability density functions are expressed, respectively, by

$$F(t|\boldsymbol{\gamma}) = 1-\left[1-G(t|\boldsymbol{\zeta})^{a}\right]^{b},$$

and

$$f(t|\boldsymbol{\gamma}) = abg(t|\boldsymbol{\zeta})G(t|\boldsymbol{\zeta})^{a-1}\left[1-G(t|\boldsymbol{\zeta})^{a}\right]^{b-1}.$$

The survival function is then given by

$$S(t|\boldsymbol{\gamma}) = \left[1-G(t|\boldsymbol{\zeta})^{a}\right]^{b}.$$

Using the relationship between the hazard, density, and survival functions shown in (2.1), we have that the hazard function associated with the Kum-G distribution is expressed by

$$h(t|\boldsymbol{\gamma}) = \frac{abg(t|\boldsymbol{\zeta})G(t|\boldsymbol{\zeta})^{a-1}}{1 - G(t|\boldsymbol{\zeta})^{a}}.$$

We can write the cumulative hazard function as

$$H(t|\boldsymbol{\gamma}) = -b\log\{1 - G(t|\boldsymbol{\zeta})^{a}\}, \tag{3.4}$$

and, using the Equations (2.6) and (3.4), the odds function is expressed as

$$R(t|\boldsymbol{\gamma}) = \exp\left[-b\log\{1 - G(t|\boldsymbol{\zeta})^{a}\}\right] - 1.$$

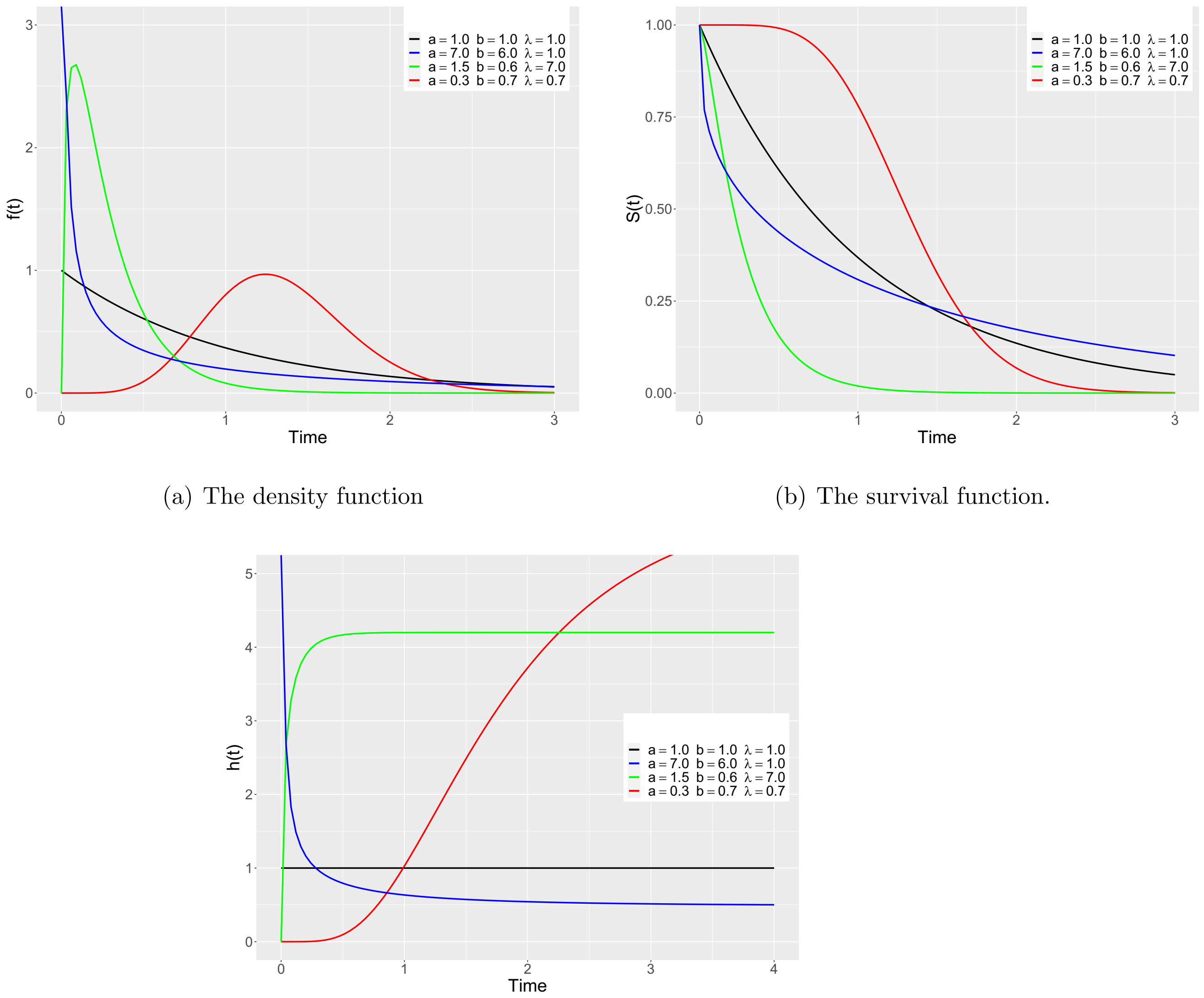
Many models can be built using (3.4). Cordeiro and de Castro (2011) presented the following special cases of the Kum-G family of distributions: the Kum-normal, Kum-Weibull, Kum-Gamma, Kum-Gumbel, and Kum-inverse Gaussian. In the next subsection, we presented the Kum-Exponential, Kum-Weibull, Kum-Gamma, Kum-Log-logistic, and Kum-Log-normal distributions used in this work to denoted the failure time.

3.1.1 Kumaraswamy-Exponential distribution

The Kumaraswamy-Exponential (Kum-Exp) distribution was presented in Cordeiro et al. (2010), as a particular case of the Kumaraswamy-Weibull and Kumaraswamy-Gamma distributions. Thus, when consider the cumulative distribution function of the exponential distribution $G(t|\boldsymbol{\zeta}) = 1 - \exp(-t\lambda)$ we can be express the Kum-Exp. Let T be a non-negative random variable denoting the time until an occurrence of an event of interest. So, we can write $T \sim$ Kum-Exp(a, b, λ) where a, b are shape parameters and λ is a scale parameter, and let $\boldsymbol{\gamma} = (a, b, \boldsymbol{\zeta})^{\top}$. Therefore, the cumulative distribution, survival, hazard, cumulative hazard, and odds functions can be express by

$$\begin{aligned}
F(t|\boldsymbol{\gamma}) &= 1 - \left[1 - (1 - \exp(-t\lambda))^{a}\right]^{b}, \\
S(t|\boldsymbol{\gamma}) &= \left(1 - \{1 - \exp(-t\lambda)\}^{a}\right)^{b}, \\
h(t|\boldsymbol{\gamma}) &= \frac{ab\lambda \exp(-t\lambda)\,(1 - \exp(-t\lambda))^{a-1}}{1 - \{1 - \exp(-t\lambda)\}^{a}}, \\
H(t|\boldsymbol{\gamma}) &= -b \ln\left(1 - \{1 - \exp(-t\lambda)\}^{a}\right), \\
R(t|\boldsymbol{\gamma}) &= \exp\left[-b \ln\left(1 - \{1 - \exp(-t\lambda)\}^{a}\right)\right] - 1.
\end{aligned}$$

The Kumaraswamy-Exp distribution was used in several studies. For instance, Adepoju and Chukwu (2015) to fit survival models to three different data sets using the maximum likelihood approach. In D'Andrea et al. (2018), this distribution was used to model survival data with a cure fraction, and Chacko and Mohan (2017) applied this distribution to model survival data subjected to type II censoring. In Figure 3.1 we present some shapes of the density, survival, and hazard function of the Kumaraswamy-Exp distribution.



(a) The density function

(b) The survival function.

(c) The hazard function.

Figure 3.1: The density, survival, and hazard functions of the Kumaraswamy-Exponential distribution.

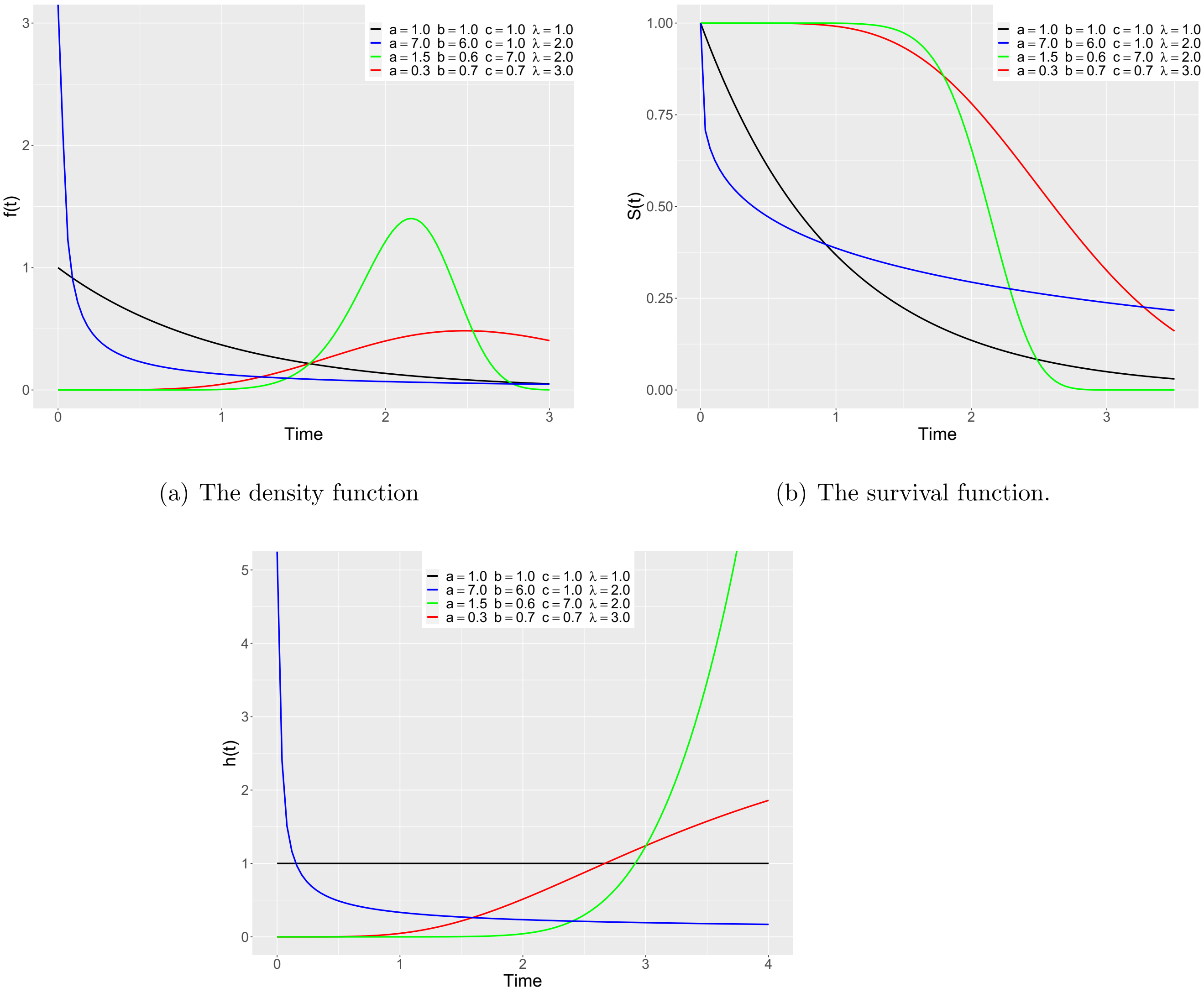
3.1.2 Kumaraswamy-Weibull distribution

In this work, we used the c.d.f $G(t|\boldsymbol{\zeta}) = 1 - \exp\left[-\left(\dfrac{t}{\lambda}\right)^c\right]$ of the Weibull distribution with $\boldsymbol{\zeta} = (c, \lambda)$, shape parameter $c > 0$ and scale parameter $\lambda > 0$. The Kumaraswamy-Weibull (Kum-W) distribution was presented in Cordeiro et al. (2010). The cumulative distribution, survival, hazard, cumulative hazard, and odds functions are given by

$$
\begin{aligned}
F(t|\boldsymbol{\gamma}) &= 1 - \left[1 - \left(1 - \exp\left[-\left(\frac{t}{\lambda}\right)^c\right]\right)^a\right]^b, \\
S(t|\boldsymbol{\gamma}) &= \left(1 - \left\{1 - \exp\left[-\left(\frac{t}{\lambda}\right)^c\right]\right\}^a\right)^b, \\
h(t|\boldsymbol{\gamma}) &= \frac{\frac{abc}{\lambda}\left(\frac{t}{\lambda}\right)^{c-1}\exp\left[-\left(\frac{t}{\lambda}\right)^c\right]\left(1 - \exp\left[-\left(\frac{t}{\lambda}\right)^c\right]\right)^{a-1}}{1 - \left\{1 - \exp\left[-\left(\frac{t}{\lambda}\right)^c\right]\right\}^a}, \\
H(t|\boldsymbol{\gamma}) &= -b\ln\left(1 - \left\{1 - \exp\left[-\left(\frac{t}{\lambda}\right)^c\right]\right\}^a\right), \\
R(t|\boldsymbol{\gamma}) &= \exp\left[-b\ln\left(1 - \left\{1 - \exp\left[-\left(\frac{t}{\lambda}\right)^c\right]\right\}^a\right)\right] - 1.
\end{aligned}
$$

In this case we used the notation $T \sim$ Kum-W(a, b, c, λ), where $a, b, c > 0$ are shape parameters, and $\lambda > 0$ is scale parameter.

As shown in Cordeiro et al. (2010) the Kum-W distribution includes in particular cases some important distributions in the context of survival analysis, such as Kum-exponential, Kum-Rayleigh, Exponentiated Weibull, Exponentiated Rayleigh, Exponentiated exponential, Weibull, Rayleigh, Exponential. In Figure 3.2 was presented some shapes of the density, survival, and hazard functions of the Kum-W distribution, we can see that this distribution is widely flexible.



(a) The density function

(b) The survival function.

(c) The hazard function.

Figure 3.2: The density, survival, and hazard functions of the Kumaraswamy-Weibull distribution.

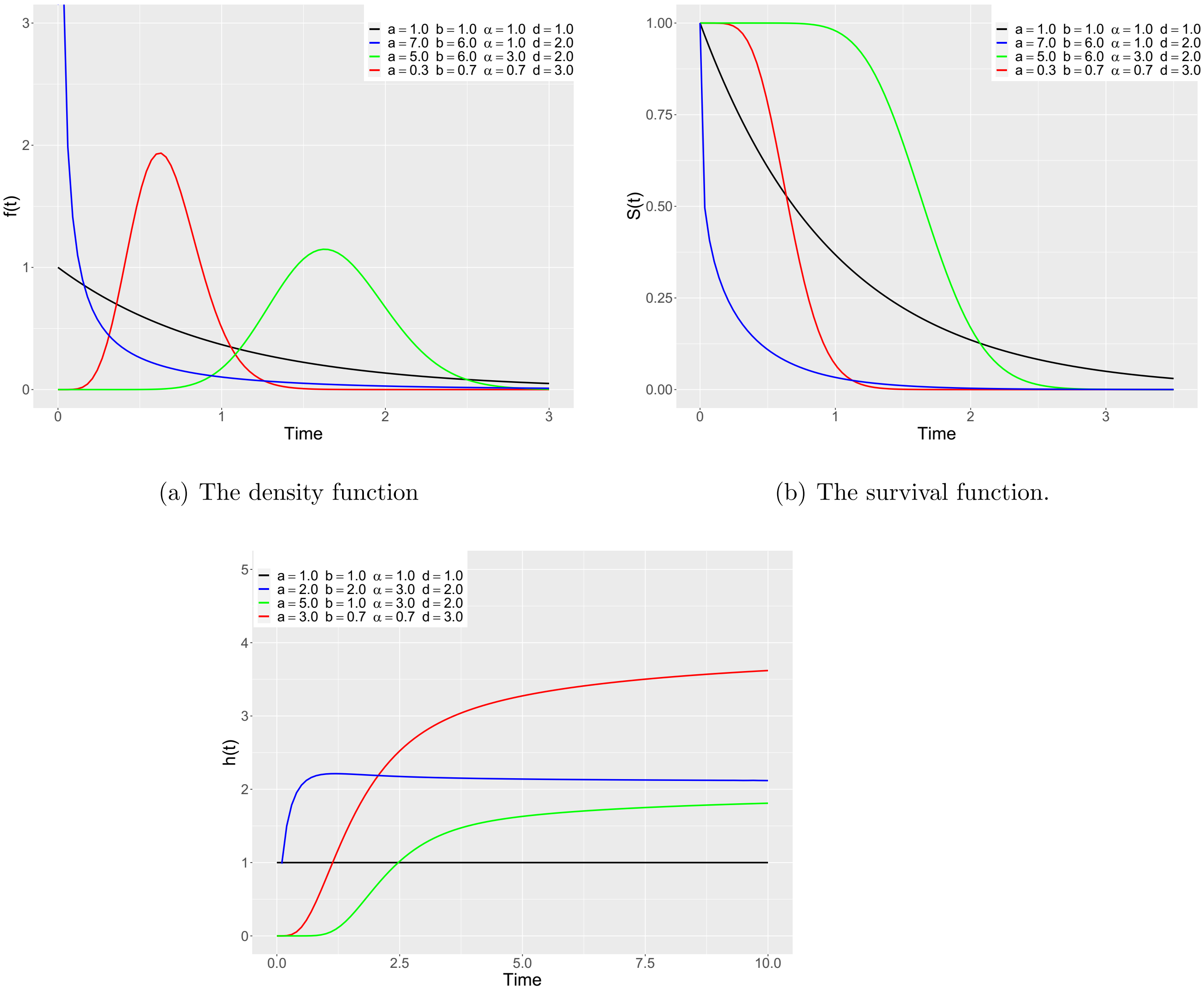
3.1.3 Kumaraswamy-Gamma distribution

Another distribution considered in this work is the Kumaraswamy-Gamma (Kum-GA) distribution, introduced in Cordeiro and de Castro (2011). Considering the cumulative distribution function $G(t|\boldsymbol{\zeta}) = \dfrac{\Gamma_{td}(\alpha)}{\Gamma(\alpha)}$ of the gamma distribution, where $\Gamma(\cdot)$ is the gamma function and $\Gamma_z = \displaystyle\int_0^z t^{(\alpha-1)}\mathrm{e}^{-t}dt$. The cumulative distribution, the survival, the hazard, the cumulative hazard, and the odds functions the Kum-GA distribution can be written by

$$
\begin{aligned}
F(t|\boldsymbol{\gamma}) &= 1 - \left[1 - \left(\frac{\Gamma_{td}(\alpha)}{\Gamma(\alpha)}\right)^a\right]^b,\\
S(t|\boldsymbol{\gamma}) &= \left[1 - \left(\frac{\Gamma_{td}(\alpha)}{\Gamma(\alpha)}\right)^a\right]^b,\\
h(t|\boldsymbol{\gamma}) &= \frac{ab\left(\frac{d^\alpha t^{\alpha-1}\exp\{-dt\}}{\Gamma(\alpha)}\right)\left(\frac{\Gamma_{td}(\alpha)}{\Gamma(\alpha)}\right)^{a-1}}{1 - \left(\frac{\Gamma_{td}(\alpha)}{\Gamma(\alpha)}\right)^a},\\
H(t|\boldsymbol{\gamma}) &= -b\log\left[1 - \left(\frac{\Gamma_{td}(\alpha)}{\Gamma(\alpha)}\right)^a\right],\\
R(t|\boldsymbol{\gamma}) &= \exp\left(-b\log\left[1 - \left(\frac{\Gamma_{td}(\alpha)}{\Gamma(\alpha)}\right)^a\right]\right) - 1.
\end{aligned}
$$

Thus, we can write $T \sim$ Kum-GA(a, b, α, d) where $a, b, \alpha > 0$ are shape parameters and d is a inverse scale parameter. It is important to mention, according to Cordeiro et al. (2010), the Kum-GA distribution has the Kum-Exp and Exponential distributions as a particular case.

Application using the Kumaraswamy-Gamma distribution can be found in De Pascoa et al. (2011) where both the maximum likelihood and Bayesian approaches are adopted for estimating the model parameters. In Figure 3.3 we presented the density, survival, and hazard functions of the Kumaraswamy-Gamma distribution. Therefore, can be seen that just like the Kum-Exp and Kum-W this distribution has many shapes, then this distribution is very flexible.



(a) The density function

(b) The survival function.

(c) The hazard function.

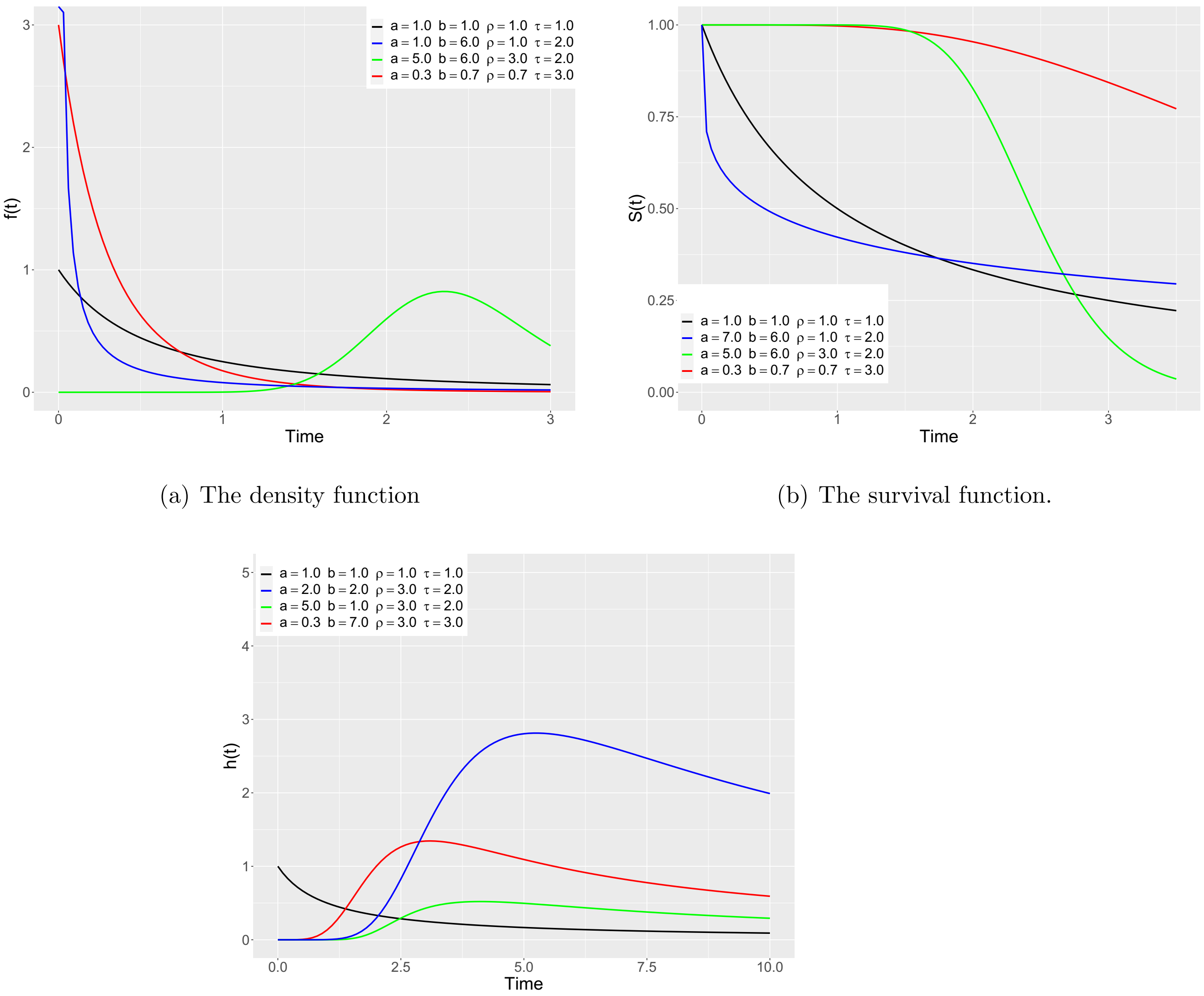
Figure 3.3: The density, survival, and hazard functions of the Kumaraswamy-Gamma distribution.

3.1.4 Kumaraswamy-Log-Logistic distribution

The Kumaraswamy-Log-logistic distribution (Kum-llogis) was presented in Santana et al. (2012). This distribution is derived from the log-logistic distribution, from the point of view of survival analysis, this distribution is very attractive, because according to Collett (2015), this distribution belongs to the class of PO and AFT models, respectively. The cumulative distribution, survival, hazard, cumulative hazard, and odds functions for the Kum-llogis distribution can be written by

$$
\begin{aligned}
F(t|\boldsymbol{\gamma}) &= 1-\left[1-\left(\frac{\tau\left(\frac{t}{\rho}\right)^{\tau}}{t\left[1+\left(\frac{t}{\rho}\right)^{\tau}\right]^{2}}\right)^{a}\right]^{b},\\
S(t|\boldsymbol{\gamma}) &= \left[1-\left(\frac{\tau\left(\frac{t}{\rho}\right)^{\tau}}{t\left[1+\left(\frac{t}{\rho}\right)^{\tau}\right]^{2}}\right)^{a}\right]^{b},\\
h(t|\boldsymbol{\gamma}) &= \frac{ab\left(1-\left[\left(\frac{t}{\rho}\right)^{\tau}\right]^{-1}\right)\left(\frac{\tau\left(\frac{t}{\rho}\right)^{\tau}}{t\left[1+\left(\frac{t}{\rho}\right)^{\tau}\right]^{2}}\right)^{a-1}}{1-\left(\frac{\tau\left(\frac{t}{\rho}\right)^{\tau}}{t\left[1+\left(\frac{t}{\rho}\right)^{\tau}\right]^{2}}\right)^{a}},\\
H(t|\boldsymbol{\gamma}) &= -b\log\left[1-\left(\frac{\tau\left(\frac{t}{\rho}\right)^{\tau}}{t\left[1+\left(\frac{t}{\rho}\right)^{\tau}\right]^{2}}\right)^{a}\right],\\
R(t|\boldsymbol{\gamma}) &= \exp\left(-b\log\left[1-\left(\frac{\tau\left(\frac{t}{\rho}\right)^{\tau}}{t\left[1+\left(\frac{t}{\rho}\right)^{\tau}\right]^{2}}\right)^{a}\right]\right)-1.
\end{aligned}
$$

In this case, we use notation for $T \sim$ Kum-llogis(a, b, ρ, τ) when, $a, b, \rho > 0$ is the scale parameter and $\tau > 0$ is a shape parameter. Santana et al. (2012) discuss the method of maximum likelihood to estimate the model parameters and two real data sets was used in the Kum-llogis distribution. Lastly, in Figure 3.4, we can note that the Kumaraswamy-Log-Logistic distribution is very flexible because has many shapes of the density, survival, and hazard functions.



(a) The density function

(b) The survival function.

(c) The hazard function.

Figure 3.4: The density, survival, and hazard functions of the Kumaraswamy-Log-Logistic distribution.

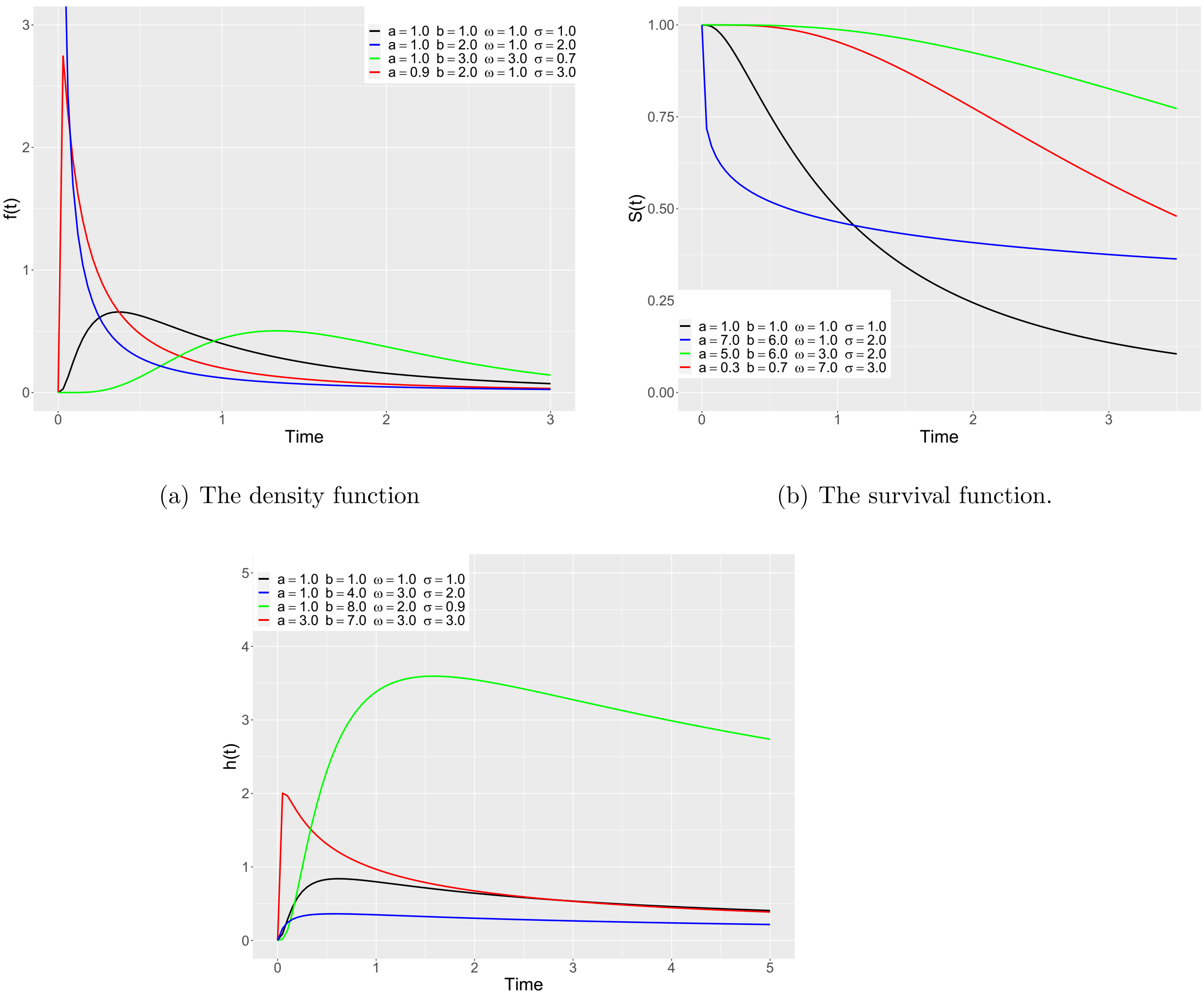
3.1.5 Kumaraswamy-Log-normal distribution

Finally, we present the Kumaraswamy-Log-normal distribution, which was introduced in Nadarajah and Rocha (2016). For this distribution we assume $G(t|\boldsymbol{\zeta}) = \left[\Phi\left(\frac{\log(t)-\omega}{\sigma}\right)\right]$, where $\Phi(\cdot)$ is the standard normal distribution. It is important to highlight that, we consider the parameterization $\omega = \log(\mu)$. Therefore, the cumulative distribution, survival, hazard, cumulative hazard, and odds functions of the Kum-lnorm can be express by

$$
\begin{aligned}
F(t|\boldsymbol{\gamma}) &= 1 - \left[1 - \left[\Phi\left(\frac{\log(t)-\omega}{\sigma}\right)\right]^{a}\right]^{b}, \\
S(t|\boldsymbol{\gamma}) &= \left[1 - \left[\Phi\left(\frac{\log(t)-\omega}{\sigma}\right)\right]^{a}\right]^{b}, \\
h(t|\boldsymbol{\gamma}) &= \frac{ab\left[e^{-\left(\frac{(\log(t)-\omega^2}{2\sigma^2}\right)}\frac{1}{t\sigma\sqrt{2\pi}}\right]\left[\Phi\left(\frac{\log(t)-\omega}{\sigma}\right)\right]^{a-1}}{1-\left[\Phi\left(\frac{\log(t)-\omega}{\sigma}\right)\right]^{a}}, \\
H(t|\boldsymbol{\gamma}) &= -b\log\left(1 - \left[\Phi\left(\frac{\log(t)-\omega}{\sigma}\right)\right]^{a}\right), \\
R(t|\boldsymbol{\gamma}) &= \exp\left[-b\log\{1 - \left[\Phi\left(\frac{\log(t)-\omega}{\sigma}\right)\right]^{a}\}\right] - 1.
\end{aligned}
$$

Then, we can write $T \sim$ Kum-Lnorm(a, b, ω, σ) where a, b, ω are shape parameters and σ is a scale parameter. Although this distribution does not have many applications in the literature, we can see in Figure 3.5, that this distribution has many forms, making this distribution very attractive to model the failure time.

Therefore, as these distributions are very flexible, that is, these distributions have many forms of density, survival, and hazard functions, we decided to consider them in the `KGsurv` package. It is important to note that many other distributions that are part of the Kumaraswamy-G distribution family can be included in the new versions of the `KGsurv` package.



(a) The density function

(b) The survival function.

(c) The hazard function.

Figure 3.5: The density, survival, and hazard functions of the Kumaraswamy-Log-normal distribution.

3.2 Regression models with Kum-G baseline distributions

In survival analysis, it is important to study the effects of covariates on the response variable, for this reason, the Cox model is the most popular in survival analysis because it is the first model that includes covariates through the hazard function. Then, we discuss an approach to including covariates information for the models implemented in the R package `KGsurv`.

To the best of our knowledge, the Kumaraswamy-G with PH, PO, and AFT families of regression models have not been considered in the literature. In this fashion, below we introduced these models.

3.2.1 Proportional hazards regression models with Kumaraswamy-G baseline distribution.

Since the introduction of proportional hazards models, many parametric distributions have been used to model the baseline hazard function, such as the Weibull, Log-normal, Gamma, Log-logistic distributions, among others. The Kum-G family of distributions arises as an attractive alternative to model the baseline hazard because it accommodates hazard functions of various shapes, adding great flexibility in the modeling. In this fashion, we can build a class of proportional hazards models based on this distribution (called Kum-G-PH).

The hazard function for the Kumaraswamy-G family of distributions is given by

$$h(t|\boldsymbol{\theta}, \mathbf{x}) = \left[\frac{abg(t|\boldsymbol{\zeta})G(t|\boldsymbol{\zeta})^{a-1}}{1 - G(t|\boldsymbol{\zeta})^{a}}\right] e^{\mathbf{x}_i^\top \boldsymbol{\beta}},$$

with $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})^\top$, where $\boldsymbol{\gamma}$ is a vector of the parameters of the Kumaraswamy-G family of distributions and $\boldsymbol{\beta}$ is a vector of the regression coefficients. Then, using (2.9), the survival function for the Kumaraswamy-G family of distributions is expressed as

$$S(t|\boldsymbol{\theta}, \mathbf{x}) = \exp\{b \log\{1 - G(t|\boldsymbol{\zeta})^{a}\} e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\}.$$

The general likelihood function of the Kum-G-PH model assumes the form

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} \left[\left[\frac{abg(y_i|\boldsymbol{\zeta})G(y_i|\boldsymbol{\zeta})^{a-1}}{1-G(y_i|\boldsymbol{\zeta})^a} \right] e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}} \right]^{\delta_i} \exp\{b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}}\}. \qquad (3.5)$$

3.2.2 Proportional odds regression models with Kumaraswamy-G baseline distributions.

Similarly to the PH model, the PO model can be build using the Kum-G family of distributions (called Kum-G-PO) using the Equations (2.6) and (2.12) the hazard function of this model can be expressed by

$$h(t|\boldsymbol{\theta},\mathbf{x}) = \frac{\exp\left[-b\log\{1-G(t|\boldsymbol{\zeta})^a\}\right] abg(t|\boldsymbol{\zeta})G(t|\boldsymbol{\zeta})^{a-1}(1-G(t|\boldsymbol{\zeta})^a)e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}}}{1+\left(\exp\left[-b\log\{1-G(t|\boldsymbol{\zeta})^a\}\right]-1\right)e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}}}.$$

The survival function of this model can be written using the Equations (2.6) and (2.13). Thus, its expression is given by

$$S(t|\boldsymbol{\theta},\mathbf{x}) = \left[1+\left(\exp\{-b\log(1-G(t|\boldsymbol{\zeta})^a)\}-1\right)e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}}\right]^{-1}.$$

From (2.7), the general likelihood function for this model can be expressed as

$$\begin{aligned} L(\boldsymbol{\theta}|D) = & \prod_{i=1}^{n} \left[\frac{\exp\left[-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\right] abg(t|\boldsymbol{\zeta})G(y_i\boldsymbol{\zeta})^{a-1}(1-G(y_i|\boldsymbol{\zeta})^a)e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}}}{1+\left(\exp\left[-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\right]-1\right)e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}}} \right]^{\delta_i} \times \\ & \left[1+\left(\exp\{-b\log(1-G(y_i|\boldsymbol{\zeta})^a)\}-1\right)e^{\mathbf{x}_i^{\top}\boldsymbol{\beta}}\right]^{-1}. \qquad (3.6) \end{aligned}$$

3.2.3 Accelerated failure time regression models with Kumaraswamy-G baseline distribution

Finally, we present the AFT regression model considering the Kum-G family of distributions, this model was named Kum-G-AFT.

The hazard function of the Kum-G-AFT model, can be expressed by

$$h(t|\boldsymbol{\theta},\mathbf{x}) = \frac{1}{\exp\{\mathbf{x}_i^{\top}\boldsymbol{\beta}\}} \frac{abg\left(\frac{t}{\exp\{\mathbf{x}_i^{\top}\boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right) G\left(\frac{t}{\exp\{\mathbf{x}_i^{\top}\boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a-1}}{1-G\left(\frac{t}{\exp\{\mathbf{x}_i^{\top}\boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a}}.$$

The survival function of this model, it is given by

$$S(t|\boldsymbol{\theta},\mathbf{x}) = \left[1-G\left(\frac{t}{\exp\{\mathbf{x}_i^{\top}\boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a}\right]^{b-1}.$$

Using the (2.17), the general likelihood function for the Kum-GAFT model is given by

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} \left[\frac{1}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} \frac{abg\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right) G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a-1}}{1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a}} \right]^{\delta_i} \left[1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a}\right]^{b-1}. \tag{3.7}$$

3.3 Inferential procedures

For the estimation of the parameters of the Kum-G-PH, Kum-G-PO, and Kum-G-AFT models, consider the following penalized log-likelihood function of the models can be expressed by

$$l_p(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}|D)) + \log(p(a, b)), \tag{3.8}$$

where $p(a, b) = f(a|\kappa, \kappa)f(b|\kappa, \kappa)$, $f(\cdot|\kappa, \kappa)$ corresponds the joint distribution of two independent random variables following gamma distributions with mean equal 1 and variance equal $1/\kappa$, and $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})^\top$ is a vector of baseline parameters and regression coefficients. The inclusion of the penalty function in (3.8) is needed to circumvent identifiability problems that might arise in the model fitting process.

The first derivatives of the log-likelihood function with respect to the $\boldsymbol{\theta}$ are presented in Appendix A. The maximum likelihood estimators and observed Fisher information matrix for the Kum-G-PH, Kum-G-PO, and Kum-G-AFT models do not have closed forms, and need to be obtained numerically, then we also used the `rstan::optimizing` function. Besides, it is important to note that the derivatives presented in Appendix A can be calculated numerically considering `rstan::optimizing` function available in R.

It is well known that, under mild regularity conditions, see Cox and Hinkley (1979), the ML estimator $\hat{\boldsymbol{\theta}}$ is consistent and follows a normal asymptotic joint distribution with an asymptotic mean $\boldsymbol{\theta}$, and an asymptotic covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ that can be obtained from the corresponding expected Fisher information matrix. So, we have $n \to \infty$

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to N_{m+p}\left(\mathbf{0}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right),$$

where m is the size of the vector $\boldsymbol{\gamma}$ and $\mathbf{0}$ is a $(m + p) \times 1$ vector.

The asymptotic confidence interval $(1-\alpha)100\%$ for θ_i can be express by

$$IC[\theta_i; \times 100(1-\alpha)\%] = \hat{\theta}_i \pm z_{\alpha/2}\sqrt{\mathcal{F}_{ii}^{-1}(\hat{\boldsymbol{\theta}})},$$

where $i = 1, 2, \ldots, m+p$.

3.4 Model selection

In this work, we consider three statistics that are used to verify the fit of the models under the frequentist approach. The first statistic is called the Akaike information criterion (AIC) (Akaike, 1998), this measure can be expressed as

$$AIC = -2l(\boldsymbol{\zeta}|D) + 2q,$$

where $l(\boldsymbol{\zeta}|D)$ is the log-likelihood function and q is the number of parameters in model.

The second statistic considered is called Bayesian Information Criteria (BIC), therefore it is given by

$$BIC = -2l(\boldsymbol{\zeta}|D) + q\log(n),$$

with n is the sample size.

The third statistic used in this work is the mean square error (MSE). This statistic can be written as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2,$$

where Y_i is the vector of observed values of the variable being predicted, with $\hat{Y_i}$ being the predicted values. Thus, the models that have a lower value of AIC, BIC, and MSE are the most adequate model for the data set used. In this work, the MSE is Residual standard error calculated using the linear regression model, where Y represents the estimated survival function calculated by the Cox-Snell residuals of the Kum-G regression models and X represents the survival function calculated by the method of Kaplan-Meyer.

The search for the best fitted models, taking into account the criteria presented above, can be carried out with the aid of the Pareto set of solutions. An optimal Pareto set is a set of solutions that are not dominated by each other (Konak et al., 2006). A non-dominated solution is defined for this problem as a solution that is never worse than the others in both objectives simultaneously. Figure 3.6 represents the bests solutions using the Pareto set solutions. The red points in the graph are the optimal points of Pareto, Although other criteria can be used, in this work we shall consider the AIC and the MSE to obtain the set of Pareto solutions.

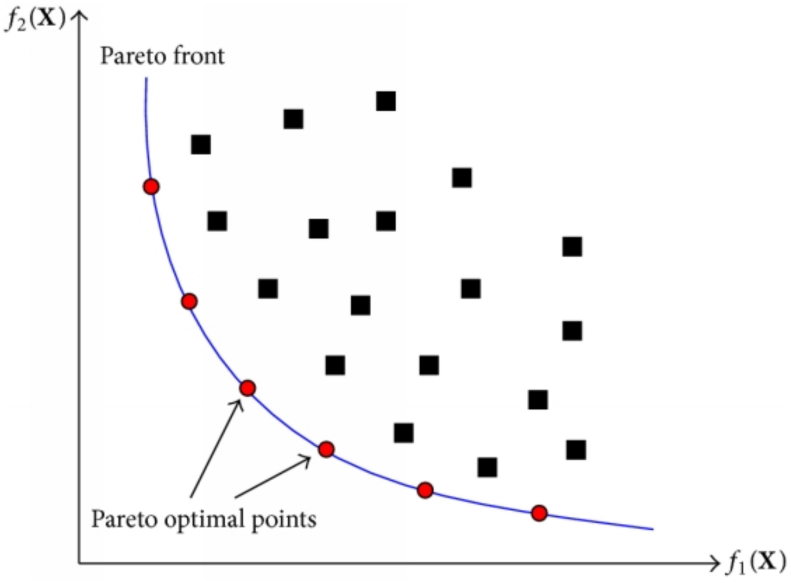


Figure 3.6: An example of a graph using the set of Pareto solutions.

Source: (Wang et al., 2015, p. 4).

3.5 R package KGsurv

In this dissertation, we propose a new package in R language, called `KGsurv`. This package was created to fit the proportional hazards, proportional odds, and accelerated failure time models considering different distributions belonging to the Kum-G family of distributions. The help for used the `KGsurv` package was presented in Appendix D. However, in this section, some theoretical concepts regarding the `KGsurv::kgreg`, `KGsurv::coxsnell`, `KGsurv::explore_fits` and `KGsurv::best_fits` functions was presented.

The function `KGsurv::kgreg` enables one to fit the PH, the PO, and the AFT families of survival regression models with baseline distributions modeled by the Kum-Exp, Kum-

W, Kum-GA, Kum-llogis, and Kum-lnorm distributions using the inferential procedures described in the previous sections, that is, under the maximum likelihood approach considering right-censored and under the assumption of a non-informative censoring mechanism. The score functions and observed Fisher information matrices, needed to carry out inferences on the models parameters under the ML approach, are obtained calculated numerically from the `rstan::optimizing` function.

Aiming to reduce problems of convergence of the algorithm, possibly related to identifiability issues that might arise in the Kum-G family of distributions, a penalty function was added to the log-likelihood function, then the log-likelihood function can be express (3.8). Specifically, the log-density of a gamma distribution with mean one and variance $1/\kappa$, was considered for the shape parameters a and b of the Kum-G distribution. Such a strategy has proven to reduce considerably the problems of convergence observed in early versions of the proposed package.

Regarding the fits of the models presented in the `KGsurv` package, some graphic methods can be used to check if the models are reasonable. For example, in (Colosimo and Giolo, 2006, p. 124 and 166) the Cox-Snell residuals are presented. These residuals are useful for examining the general fit of the PH, PO, and AFT models. The Cox-Snell residuals were introduced in Cox and Snell (1968) and can be express by

$$\hat{e}_i = \hat{H}(t|\boldsymbol{\beta}, \mathbf{x}_i), \tag{3.9}$$

where $H(t|\cdot)$ is the cumulative hazard function presented in (2.1). Then, using relation (2.4) and Equations (2.9), (2.13), and (2.16) can be provide the residuals of the Cox-Snell for the PH, PO, and AFT models. In this work, the `KGsurv::coxsnell` function was presented to calculate the Cox-Snell residuals of the models presented in the `KGsurv` package to assist in choosing the model with the best fit.

Another attractive function available in the `KGsurv` package is the `KGsurv::explore_fits` function. This function was developed with the aim to allow the user to fit/explore a large number of models, taking into account different choices of regression structures and baseline distributions, in combination with different values for the penalty parameter κ (by default, 1e-03, 1e-02, 1e-01, 0, 1e+00,1e+01, 1e+02, 1e+03). Therefore, the

KGsurv::explore_fit function allows one to explore up to 120 models considering all combinations of regression models, failure time distributions, and penalties considered in this work.

Finally, after saving the result of the KGsurv::explore_fits function on an object, the KGsurv::best_fits function must be used. This function provides a graph similar to that displayed in Figure 3.6, along with a table containing the best fitted models, according to the Pareto solution presented in the previous section. It is important to mention that, only the models to which the algorithms converged are chosen and the order in which the models are placed is decreasing concerning the values of the MSE statistic, that is, from the highest to the lowest. Additionally, the KGsurv::best_fits function also provides a graphic containing the plots of the Cox-Snell residuals associated with the best fits, together with the penalty parameter, the AIC, and MSE values.

In the next chapter, we present three applications using the models implemented in the KGsurv package considering three real data sets widely used in survival analysis literature.

Chapter 4

Applications

In this chapter, we use the models implemented in the R package `KGsurv` to reanalyze three real data sets that have been previously addressed in the literature. Our goal here is to compare the models fitted using the R package `KGsurv` with the Bernstein polynomial models presented in Appendix B and Panaro (2020), and available in R package `spsurv`. The three data sets involve right-censored survival data and the proportional hazards assumption being valid for the first data set, and invalid for the other ones.

4.1 Laryngeal cancer data set

In this section, we present an application using the Kum-G-PH, Kum-G-PO, and Kum-G-AFT models considering the five G distribution discussed in the previous sections to analyze the data set from a study presented in Klein and Moeschberger (2006), and available in the R package `KMsurv`. This study involved 90 male patients diagnosed with larynx cancer between the years 1970 to 1978, with follow-up until 1983. The censorship percentage of this data is approximately 44%. The response and covariates used in this application are time until death (in the month), standardized age (between 0 or 1), and stage (I, II, III, or IV). In this clinical trial, we consider Stage I as the reference group.

One of the aims of this study was to investigate whether the age and stage of cancer are related to the death of patients with laryngeal cancer. In short, in this clinical trial, the average is 79 age is 64.61 and the standard deviation is 10.79, 37% of patients are stage I, 19% stage II, 30% stage III, and 14% stage IV.

Figure C.2, shows the standardized Schoenfeld residuals along with the test of proportional hazards assumption of the Cox model for each covariate included in the model. Considering the significance level of 5% and based on Figure C.2, there is no evidence to reject the proportional hazards hypothesis. Therefore, we expect the Kum-G-PH model to provide a good fit, however, we use all models in this data set.

Table 4.1 presents the best five models provided by the `KGsurv::explore_fits` and `KGsurv::best_fits` functions for all models considering the Laryngeal cancer data set. In Table 4.1, we can see that the five models considered are Kum-lnorm-PH, Kum-llogis-PO, and Kum-lnorm-PO with penalty parameters ranging from 1 to 1000 with code equal 0, that is, the algorithm converged for these models (code different of 0 indicate that the algorithm no converged). Besides, the `KGsurv::best_fits` show the plots of the Pareto solution and Cox-Snell residuals, respectively. Then, in Figure 4.3, we present the five models chosen from the 120 models fit by the `KGsurv::explore_fits` function. Besides in Figure 4.1, we can see that, the Kum-lnorm-PH, Kum-llogis-PO, and Kum-lnorm-PO present a good fit for the data set. It can be noted that the Kum-lnorm-PH has the smallest AIC and Kum-lnorm-PO has the smallest MSE, but the Cox-Snell residuals shown in Figures 4.1 and 4.2, indicates that the Kum-llogis-PO model with a penalty equal to 1000 has the best fit, considering that the points are closer to the red line.

Table 4.1: The best fits of the Kum-G-PH, Kum-G-PO, and Kum-G-AFT models for laryngeal data set.

Family	Distribution	AIC	MSE	Penalty	code
PH	Log-normal	270.3176	0.0254	1000	0
PO	Log-logistic	271.6043	0.0212	1000	0
PO	Log-logistic	280.6618	0.0212	100	0
PO	Log-logistic	289.1522	0.0210	10	0
PO	Log-normal	296.3727	0.0166	1	0

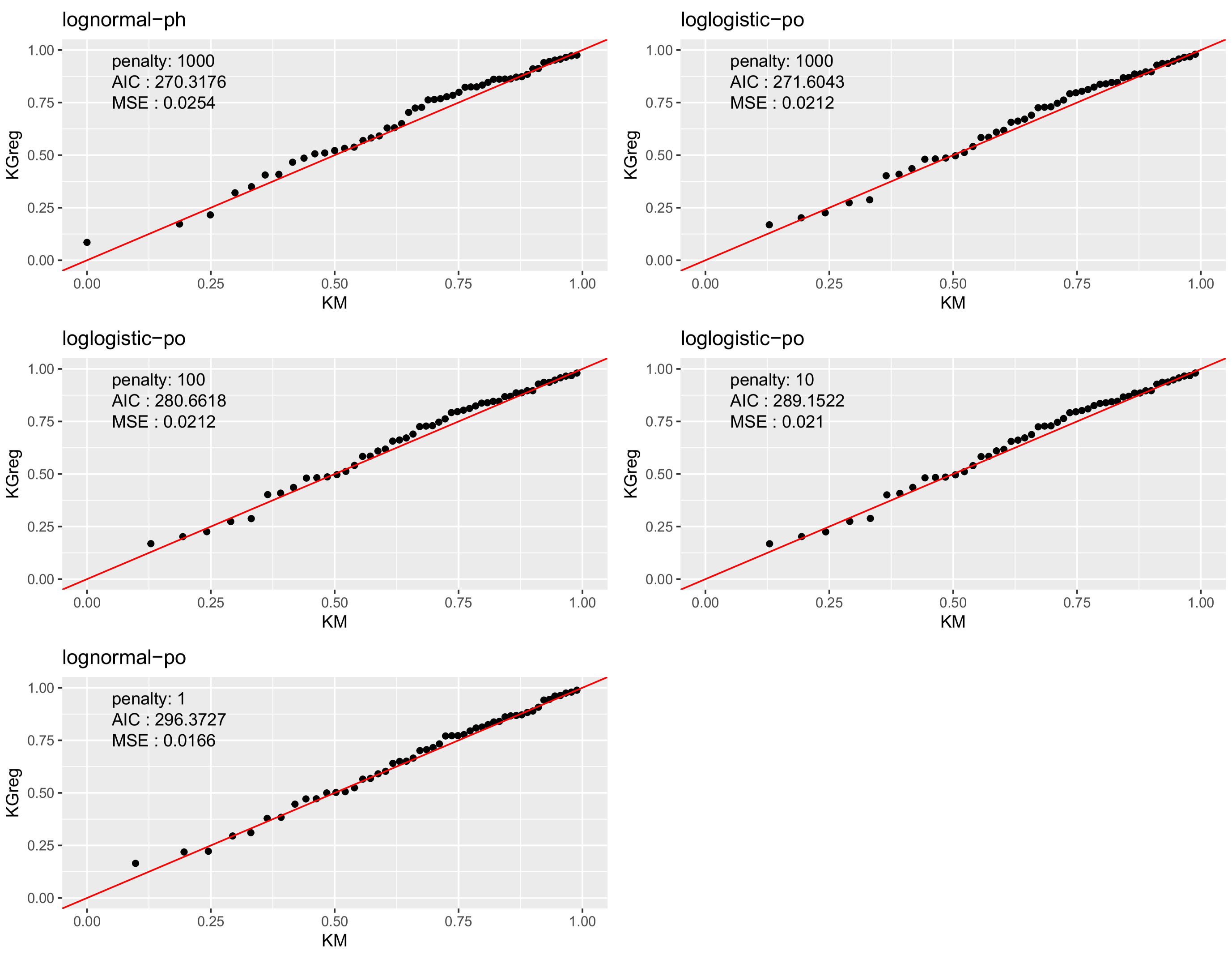


Figure 4.1: The plot of the Cox-Snell residuals for Kum-G-PH, Kum-G-PO, and Kum-G-AFT models considering the `KGsurv::best_fits` function for the laryngeal data set.

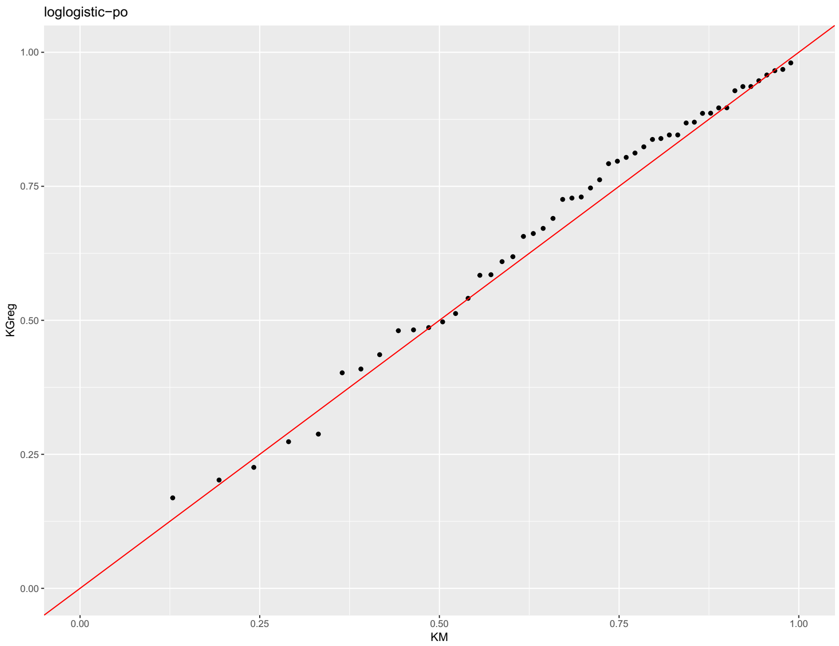


Figure 4.2: The plot of the Cox-Snell residuals for Kum-llogis-PO model for the laryngeal data set.

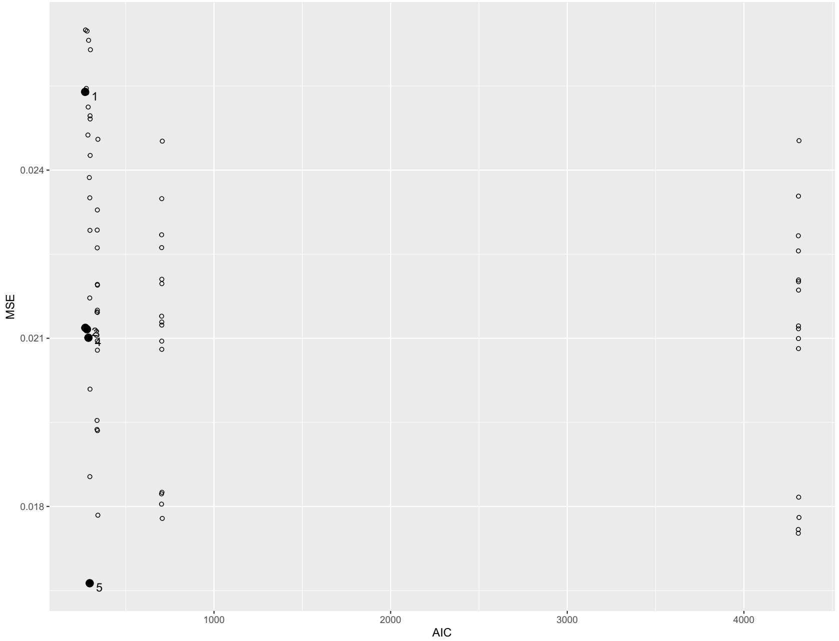


Figure 4.3: The plot of the Pareto set for the best fits of the Kum-G-PH, Kum-G-PO, Kum-G-AFT models considering the `KGsurv` package for the laryngeal data set.

The Kum-llogis-PO and the proportional odds model considering the Bernstein polynomial (BPPO) models were considered to explain the relationship between the survival time with the age of the patients and the stage of cancer (stage I was considered the reference group) and to compare whether the results are similar. Table 4.2 shows the results of the fit of the two models. As it can be observed, the Kum-llogis-PO model presents similar results when compared with the BPPO model, and the Cox-Snell residuals for Kum-llogis-PO model presented in Figure 4.2 indicate that this model presents a good fit alternative to this data set.

Table 4.2: ML for the Kum-llogis-PO and Bernstein Polynomial based Proportional Odds model for the laryngeal cancer data set.

Kum-llogis-PO model					
				CI	
	coef	exp(coef)	s.e(coef)	2.5%	97.5%
β_1 : Age	0.2240	1.2511	0.2120	−0.1915	0.6395
β_2 : Stage II	0.3657	1.4415	0.6021	−0.8145	1.5458
β_3 : Stage III	1.3199	3.7430	0.5178	0.3051	2.3348
β_4 : Stage IV	2.6598	14.2934	0.6268	1.4313	3.8883
	AIC = 271.6043		BIC = 291.6028		
Bernstein Polynomial based Proportional Odds model					
				CI	
	coef	exp(coef)	s.e(coef)	2.5%	97.5%
β_1 : Age	0.2231	1.2501	0.2084	−0.0171	0.0585
β_2 : Stage II	0.1892	1.2083	0.5895	−0.9662	1.3446
β_3 : Stage III	1.1306	3.0976	0.5045	0.1419	2.1193
β_4 : Stage IV	2.4615	11.7221	0.6294	1.2280	3.6950
	AIC = 308.0402		BIC = 343.0375		

It can also be observed that the Kum-llogis-PO model has the smallest value of AIC and BIC concerning the BPPO model, suggesting that, the Kum-llogis-PO model has the best fit for the laryngeal data set.

Regarding the interpretation of the Kum-llogis-PO model parameters, we have that the effect of the patients' age was not statistically significant. Also, the effects of patients diagnosed in stages II were not statistically significant, that is, there is no significant difference between the effects of patients in stage II concerning the reference group. However, when we consider the effect of patients diagnosed with stage III and IV disease, there is a significant difference concerning the reference group. Therefore, the estimated OR for the stage III patients is approximately 4, which means that the odds of death is about 4 times higher for patients in stage III when compared to the same age patients in stage I, whereas the estimated OR for the stage IV patients is approximately 14, which means that the odds of death is about 14 times higher for patients in stage IV when compared to the same age patients in the reference group.

In summary, the results provided by the Kum-llogis-PO model are similar to Bernstein's semi-parametric model presented in Panaro (2020), and also the plot of the Cox-Snell residuals displayed in Figure 4.2 indicates a reasonable fit for this data set.

4.2 NCCTG lung cancer data set

This study was conducted by the North Central Cancer Treatment Group and presented in Loprinzi et al. (1994), the data set is available in the R package `survival`. This clinical trial involved patients with advanced lung cancer, and the goal of the study was to determine whether patients' self-assessment could provide prognostic information complementary to the physician's assessment. The data set contains 228 patients, including 63 patients that were right-censored resulting in approximately 28% of censorship.

The response and covariates considered in clinical trial are: time until the death (in month), ph.karno: performance score (0 - bad, 100 - good), and ph.ecog: performance score as rated by the physician (0 = asymptomatic, 1 = symptomatic but completely ambulatory, 2 = in bed $< 50\%$ of the day, 3 = in bed $> 50\%$ of the day but not bed-bound, 4 = bed-bound). The performance score has an average of approximately 82 and the standard deviation is approximately 12. Lastly, the performance score as rated by the physician there are 27 % in classified in 0, 50% in 1, 21% in 2, and 1% in 4.

In Figure C.3, we can see that the assumption of proportional hazards has been violated for a ph.karno covariate considering the significance level of 5 %. However, the value is close to 0.05, also, considering higher levels of significance, this covariate accepts the assumption of proportional hazards. Therefore, we considered the Kum-G-PH models in the analysis to show the effectiveness of the `KGsurv::best_fits` function in identifying the models with the best fits for this data set.

Table 4.3 shows the best fits provided by the `KGsurv::best_fits` function. In Table 4.3 and Figure 4.6, we can see that the models Kum-Exp-AFT, Kum-Exp-PH, Kum-llogis-PO, and Kum-llogis-AFT were presented with different penalty values ranging from 0.1 to 1000. Then, based on the Table 4.3 and Figure 4.4, the Kum-Exp-AFT model presented the lowest AIC value and the Kum-llogis-PO presented the lowest MSE, however, considering the Cox-Snell residuals in Figure 4.5 the Kum-Exp-AFT model with the penalty parameter equal to 1000 presented the reasonable fits for this application.

In Table 4.4, we present the results of the Kum-Exp-AFT and the Bernstein polynomial-based Accelerated Failure time (BPAFT) models for the lung data set. We can see that

Table 4.3: The best fits of the Kum-G-PH, Kum-G-PO, and Kum-G-AFT models for lung data set.

Family	Distribution	AIC	MSE	Penalty	code
AFT	Exponential	1143.960	0.0182	1000	0
PH	Exponential	1144.821	0.0170	1000	0
PO	Log-logistic	1156.859	0.0169	100	0
AFT	Log-logistic	1172.348	0.0157	10	0
PO	Log-logistic	1212.090	0.0140	0.1	0

the β_1 for the models showed a difference and considering the significance level of 5 % the β_4 is not significant, whereas the BPAFT models are significant. However, the other parameters had similar results. Besides, we can observe that the AIC and BIC of the Kum-Exp-AFT model are inferior concerning the BPAFT model, then, for this data set these criteria suggesting that the Kum-Exp-AFT model presents the best fit.

For the interpretation of the parameters was considered the Kum-Exp-AFT model. It is important to mention that for this application, ph.ecog IV was considered as a reference group. Thus, we can notice that the ph.karno variable was not statistically significant. Besides, the ph.ecog I covariate was not statistically significant, that is, there is no difference between patients with ph.ecog I compared to patients with ph.ecog IV. The ph.ecog II covariate was statistically significant, that is, there is a difference between patients with ph.ecog II compared to patients in the reference group. Therefore, the estimated TR is 0.42, which means that the median time to death for patients with ph.ecog II has a reduced 53% concerning compared with patients with ph.ecog IV. Finally, the ph.ecog III covariate was not statistically significant, thus, there is no difference between patients with ph.ecog III concerning patients in the reference group. It is important to note that in the BPAFT model this effect is significant, then, the median time to death for patients with ph.ecog III has a reduced 68% compared with patients with ph.ecog IV.

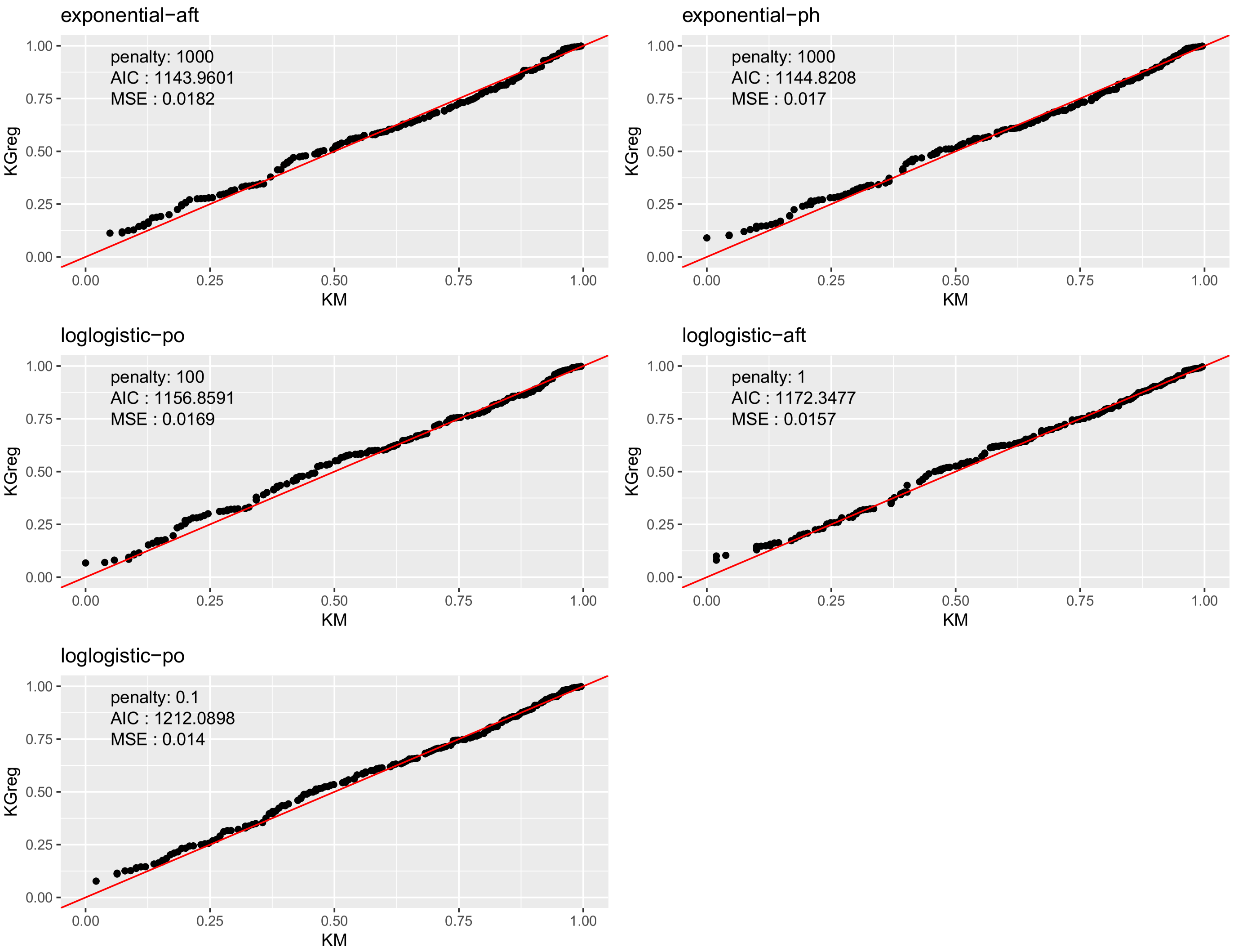


Figure 4.4: The plot of the Cox-Snell residuals for Kum-G-PH, Kum-G-PO, and Kum-G-AFT models considering the `KGsurv::best_fits` function for the lung data set.

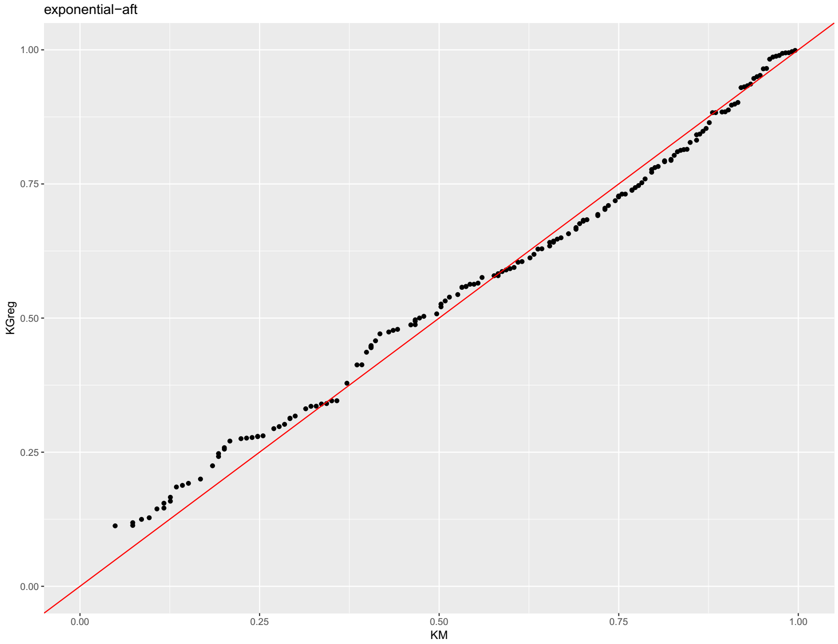


Figure 4.5: The plot of the Cox-Snell residuals for Kum-Exp-AFT model for lung data set.

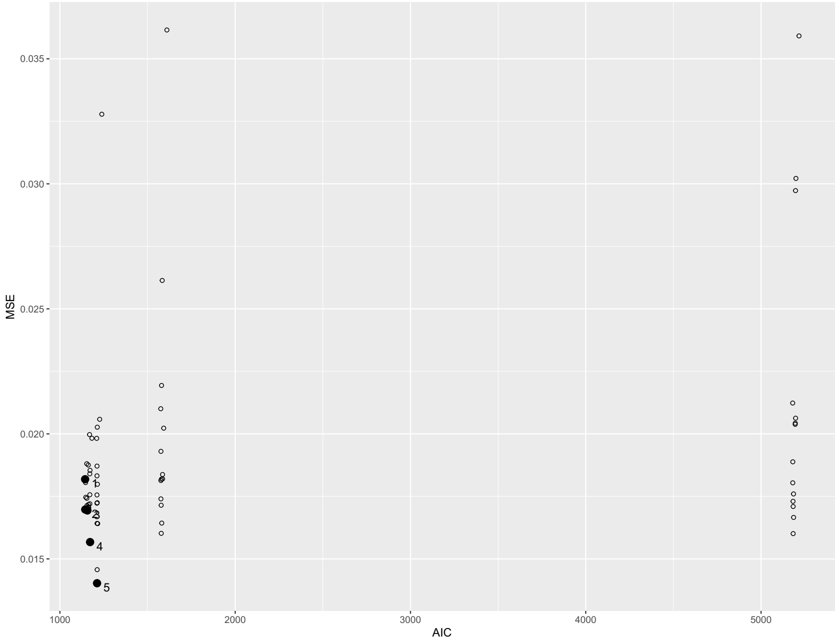


Figure 4.6: The plot of the Pareto set for the lung data set.

Briefly, in this application, we use the Kum-Exp-AFT model, which presented similar results with the results presented by the BPAFT model, except the ph.karno and ph.ecog

III covariates. However, as the Kum-Exp-AFT model presented the lowest AIC and BIC values and the Cox-Snell residuals 4.5 presented a reasonable fit, we can see that considering the Kumaraswamy-G family of distribution as a baseline is an attractive alternative for the set of lung cancer data.

Table 4.4: ML for the Kum-Exp-AFT and Bernstein Polynomial based Accelerated Failure time model for the lung data set.

Kum-Exp-AFT model					
				CI	
	coef	exp(coef)	s.e(coef)	2.5%	97.5%
β_1 : ph.karno	-0.0055	0.9945	0.0079	-0.0210	0.0101
β_2 : ph.ecog I	-0.3398	0.7119	0.1826	-0.6977	0.0182
β_3 : ph.ecog II	-0.8621	0.4223	0.2874	-1.4255	-0.2987
β_4 : ph.ecog III	-1.6448	0.1931	0.8713	-3.3525	0.0629
	AIC = 1143.96		BIC = 1167.904		
Bernstein Polynomial based Accelerated Failure time					
				CI	
	coef	exp(coef)	s.e(coef)	2.5%	97.5%
β_1 : ph.karno	0.0029	1.0029	0.0032	-0.0034	0.0093
β_2 : ph.ecog I	-0.24062	0.7861	0.1436	-0.5222	0.0409
β_3 : ph.ecog II	-0.5673	0.5671	0.1387	-0.8392	-0.2954
β_4 : ph.ecog III	-1.1321	0.3224	0.4781	-2.0692	-0.1950
	AIC = 1187.771		BIC = 1256.358		

4.3 Veterans administration data set

In this section, we use the models implemented in the proposed package to reanalyze the data set described in Prentice (1973), available in R package `survival`. This data set contains n = 137 patients who were followed up by the Veterans Administration Lung cancer study group. The censorship percentage for this clinical study is approximately 6.5%. For this clinical trial, the response and exploratory variables are time until death (in days), standardized PS: patients' performance score (from 0 to 1), and cell type: Histological type of tumor (squamous cell, small cell, adeno cell, large cell).

For the time until the death of patients, we divide the time by 30 days, so that the results of the models do not present computational problems. The covariate PS has an average of 58.57 and a standard error of 20.03. Concerning the covariate cell type, 27% are squamous cells, 34% small cells, 20% adeno cells, and 20% large cells. The interest in this application is to evaluate the effect of the PS and cell type covariates on the survival time of patients with lung cancer. For this data set, we consider the squamous cells as a reference group.

We checked whether the proportional hazards model can be applied to this data set. In Figure C.4 it shows that the basic assumption of the Cox model was violated, like the previous application, the Kum-G-PH models are not suitable for this data set, that is, the Kum-G-PO and Kum-G-AFT models are suitable for this data set. However, just like the previous application, we use `KGsurv::explore_fits` to check if all models present in the `KGsurv::kgreg` package fit for the veteran data set.

Table 4.5 and Figure 4.9 show the best models considering the Pareto set, and it can be seen that only five models were considered, the Kum-llogis-PO and Kum-llogis-AFT models considering different penalty values ranging from 1 to 1000. For this application the model chosen was the Kum-llogis-AFT model with the penalty parameter equal to 1000, this choice was based on Figure 4.7 and Figure 4.8, which shows that despite the model has not the lowest MSE and AIC value, however, the Cox-Snell residuals have a better fit for the Kum-llogis-AFT model.

In Table 4.6, it can be seen that the estimates of the regression coefficients of the

Table 4.5: The best fits of the Kum-G-PH, Kum-G-PO, and Kum-G-AFT models for Veterans administration data set.

Family	Distribution	AIC	MSE	Penalty	code
PO	Log-logistic	539.6952	0.0174	1000	0
AFT	Log-logistic	541.8205	0.0154	1000	0
AFT	Log-logistic	550.8849	0.0154	100	0
AFT	Log-logistic	559.4134	0.0152	10	0
AFT	Log-logistic	568.5501	0.0142	1	0

Kum-llogis-AFT and BPAFT models are similar. However, the AIC and BIC of the Kum-llogis-AFT model are inferior to those of the BPAFT model, that is, these results suggest that the Kum-llogis-AFT model presents a better fit. Also, the graph of the Cox-Snell residuals of this model presented in Figure 4.8 shows that the Kum-llogis-AFT has a reasonable fit for the Veterans data set. Therefore, for the analysis of the results for the Veterans data set, we consider the fit of the Kum-llogis-AFT model.

The covariate PS was statistically significant, then, if we increase a unit in the PS, we have that the median time to death has increased by 2 times. Besides, the small cell and adeno cell types were also statistically significant, that is, the median time to death of these cells is different compared to the reference cell. For the small cell, we have that the estimated TR is 0.47, this implies that the median time to death has a reduction of 53% concerning patients with squamous cells. For the adeno cell, we have an estimated TR of 0.46, which means that the median time to death has a 54% reduction concerning patients in the reference group. Finally, the large cell was not statistically significant, this implies that the median time to death does not differ from patients with a large cell compared to the squamous cell.

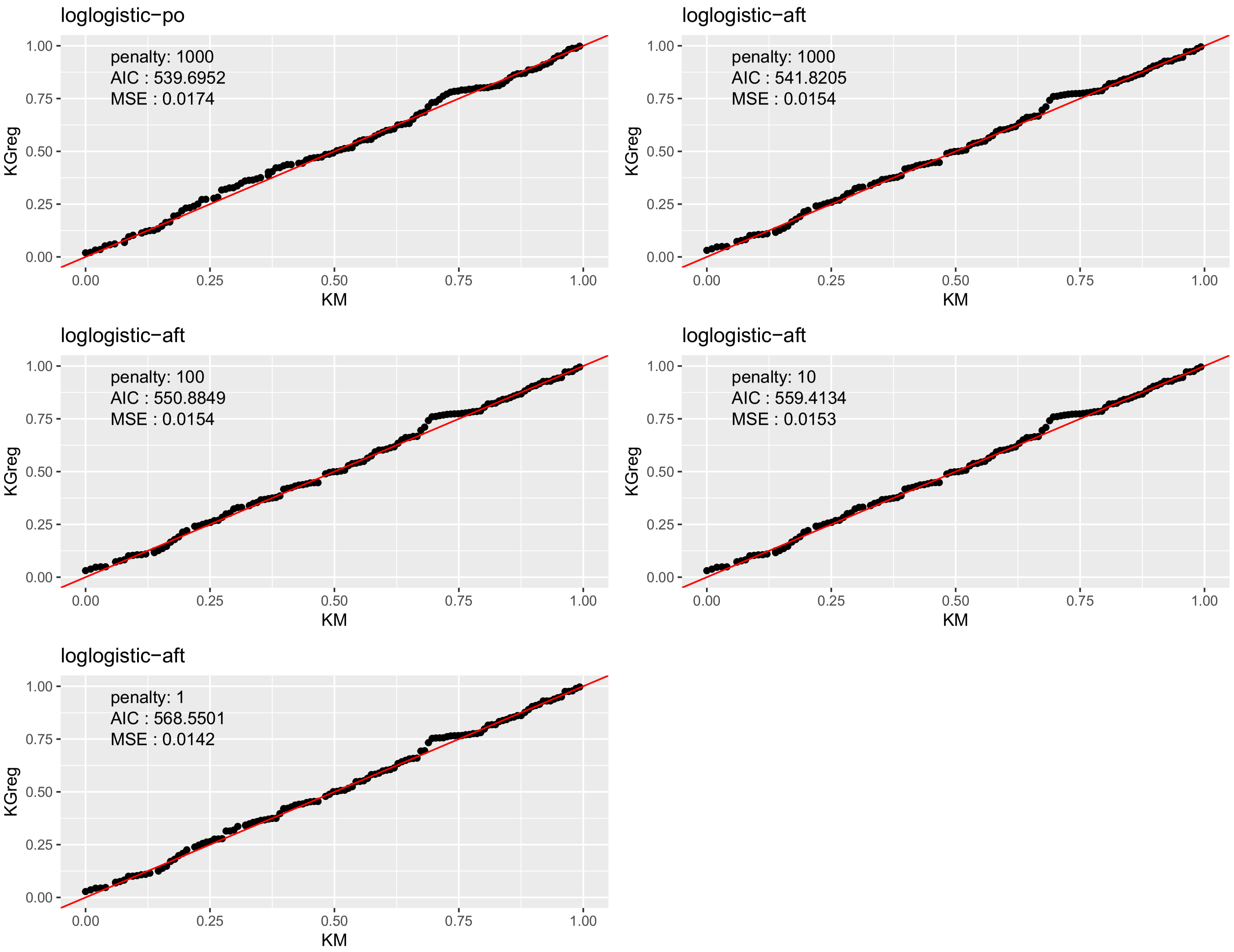


Figure 4.7: The plot of the Cox-Snell residuals for Kum-G-PH, Kum-G-PO, and Kum-G-AFT models considering the `KGsurv::best_fits` function for the Veterans data set.

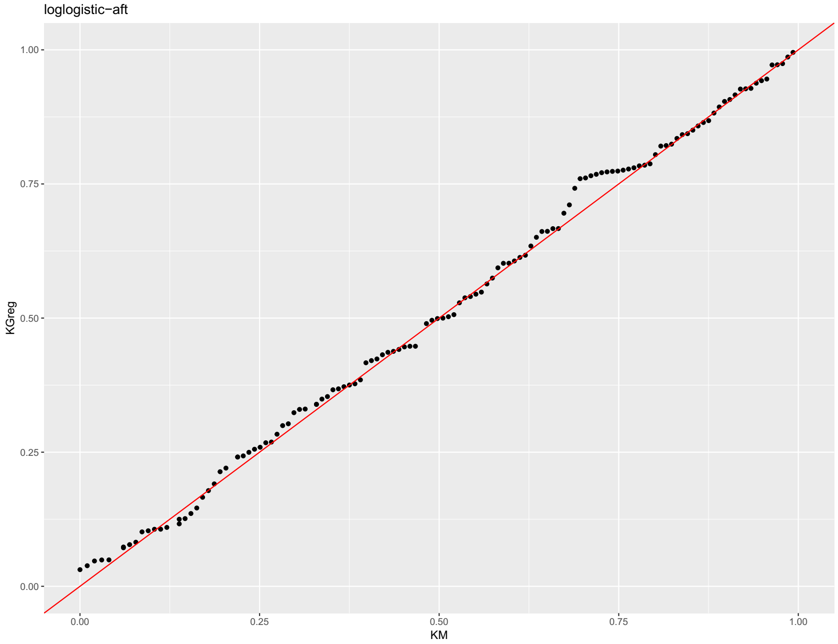


Figure 4.8: The plot of the Cox-Snell residuals for Kum-llogis-AFT model for the Veterans data set.

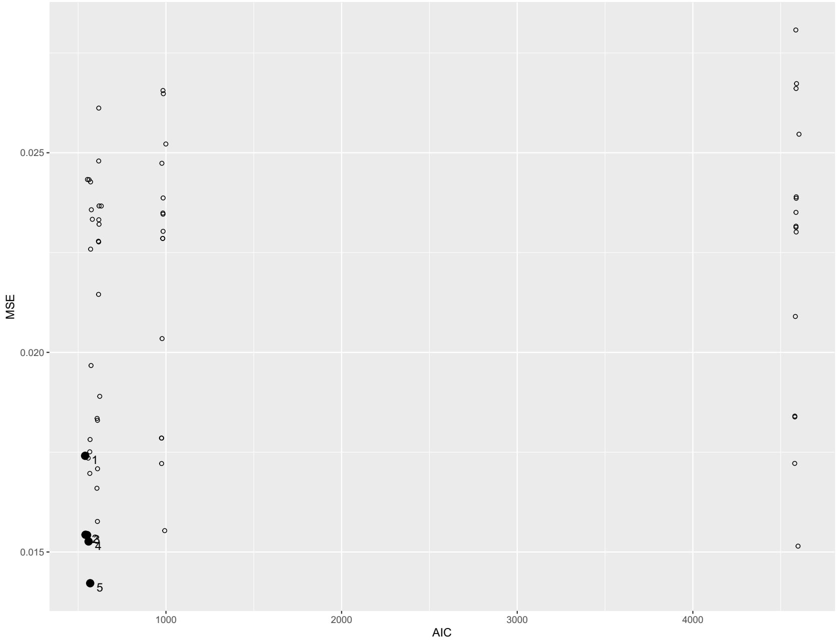


Figure 4.9: The plot of the Pareto set for the best fits of the Kum-G-PH, Kum-G-PO, Kum-G-AFT models considering the `KGsurv` package for the Veterans data set.

In short, the Kum-Exp, Kum-Weibull, Kum-Gamma, Kum-lnorm, and Kum-llogis models were used as baseline distributions for the regression models PH, PO, and AFT to analyze the Veterans data set. Based on Tables 4.5 and 4.6 and Figures 4.7, 4.8, and 4.9, we can see that the Kum-llogis model presented is one possibility for fit to this data set.

Table 4.6: ML for the Kum-llogis-AFT and Bernstein Polynomial based Accelerated Failure time model for the Veteran data set.

Kum-llogis-AFT model					
				CI	
	coef	exp(coef)	s.e(coef)	2.5%	97.5%
β_1 : PS	0.7006	2.0150	0.0823	0.5393	0.8618
β_2 : Small cell	−0.7455	0.4745	0.2346	−1.2053	-0.2857
β_3 : Adeno cell	−0.7579	0.4686	0.2562	−1.2601	-0.2556
β_4 : Large cell	−0.0921	0.9120	0.2556	−0.5931	0.4089
	AIC = 541.8205		BIC = 565.1803		
Bernstein Polynomial based Accelerated Failure time model					
				CI	
	coef	exp(coef)	s.e(coef)	2.5%	97.5%
β_1 : PS	0.6849	1.9836	0.0753	0.5374	0.8324
β_2 : Small cell	−0.6926	0.5003	0.2448	−1.1723	-0.2129
β_3 : Adeno cell	−0.8694	0.4192	0.2487	−1.3568	-0.3819
β_4 : Large cell	−0.1252	0.8823	0.2634	−0.6415	0.3911
	AIC = 580.5850		BIC = 627.3047		

Chapter 5

Conclusions

In this dissertation, we developed an R package called KGsurv, based on the Stan software, that allows one to fit the PH, PO, and AFT survival regression models, considering the Kumaraswamy-G family of distributions as the baseline distribution with the following choices for G: Exponential, Weibull, Gamma, Log-logistics, and Log-normal distributions. Inferences are carried out via the maximum likelihood approach, under the assumption of right-censored survival data subjected to non-informative censoring.

Regarding the results of the PH, PO, and AFT models considering the Kumaraswamy-G family of distributions as a baseline distribution, it is important to mention that some algorithms for the models presented convergence problems, possibly related to the question of the identifiability of the models. One way to get around these possible problems is to use a penalty in the likelihood function. Then, in this work, we use a penalty in the likelihood function of the model considering the Gamma distribution using the penalty parameter ranging from (1e-03, 1e-02, 1e-01, 0, 1e + 00,1e+ 01, 1e + 02, 1e + 03).

The usefulness of the KGsurv package is illustrated through the analysis of three real data sets that have been previously addressed in the literature. For comparison purposes, the models implemented in the proposed package are compared with the semi-parametric regression models with baseline modeled by Bernstein polynomials introduced by Panaro (2020), and available in the R package spsurv. For the applications considering the laryngeal and Veterans data sets, the fits of the Kum-G model are similar compared to

Bernstein's Polynomial model. However, for the lung data set the covariates ph.karno and ph.ecog III presented different fits for the Kum-G and the Bernstein Polynomial models, whereas that the other covariates are similar in the two models. The model selection criteria suggest that the models presented in the `KGsurv` package have the best fits compared to the models presented in the `spsurv` package. Therefore, considering the results presented in the three applications, we hope that the `KGsurv` package presents an attractive alternative to survival data with the right-censoring.

The next step is to carry out a simulation study to compare the fits of the Bernstein Polynomial models and the Kum-G models and to verify the possible convergence problem of the models presented in the `KGsurv` package. Also, another possible extension is to include a Bayesian inference approach in the `KGsurv` package and consider left-censored and interval-censored considering the informative censoring mechanism. Other examples of possible extensions are to use the cure fraction, frailty, and crossing survival curves models.

Appendix

Appendix A: The Kum-G-PH, Kum-G-PO, and Kum-G-AFT

In this chapter, the log-likelihood functions and the Score functions of the Kum-G-PH, Kum-G-PO, and Kum-G-AFT models are presented. It is important to highlight that the results are described in a generic way, in such a way that any choice of distribution of G can be used. Thus, using the Equation (3.5) the log-likelihood of the Kum-GPH models can be express by

<u>The Kum-G-PH model</u>:

$$\begin{aligned} l(\boldsymbol{\theta}|D) = \sum_{i=1}^{n} & \delta_i[\log(a) + \log(b) + \log[g(y_i|\boldsymbol{\zeta})] + (a-1)\log[G(y_i|\boldsymbol{\zeta})] + \mathbf{x}_i^\top \boldsymbol{\beta} \\ & - \log\left(1 - G(y_i|\boldsymbol{\zeta})^a\right)] + b\log\{1 - G(y_i|\boldsymbol{\zeta})^a\}e^{\mathbf{x}_i^\top \boldsymbol{\beta}} + 2(\kappa\log(\kappa) - \log(\Gamma(\kappa)) \\ & + (\kappa-1)(\log(b) + \log(a)) - \kappa(b+a). \end{aligned}$$

The elements of the score vector of the Kum-G-PH models is given by

$$\frac{\partial l(\boldsymbol{\theta}|D)}{\partial a} = \sum_{i=1}^{n} \frac{\delta_i}{a} + \delta_i\log[G(y_i|\boldsymbol{\zeta})] + \frac{\delta_i G(y_i|\boldsymbol{\zeta})^a \log[G(y_i|\boldsymbol{\zeta})]}{1 - G(y_i|\boldsymbol{\zeta})^a} - \frac{be^{\mathbf{x}_i^\top \boldsymbol{\beta}} G(y_i|\boldsymbol{\zeta})^a \log[G(y_i|\boldsymbol{\zeta})]}{1 - G(y_i|\boldsymbol{\zeta})^a} + \frac{\kappa-1}{a} - \kappa.$$

$$\frac{\partial l(\boldsymbol{\theta}|D)}{\partial b} = \sum_{i=1}^{n} \frac{\delta_i}{b} + b\log\{1 - G(y_i|\boldsymbol{\zeta})^a\}e^{\mathbf{x}_i^\top \boldsymbol{\beta}} + \frac{\kappa-1}{b} - \kappa.$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta}|D)}{\partial \zeta_k} = \sum_{i=1}^{n} & \delta_i \frac{1}{g(y_i|\boldsymbol{\zeta})} \frac{\partial g(y_i|\boldsymbol{\zeta})}{\partial \zeta_k} + \delta_i \frac{1}{G(y_i|\boldsymbol{\zeta})} \frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial \zeta_k} + \delta_i \frac{aG(y_i|\boldsymbol{\zeta})^{a-1}}{1 - G(y_i|\boldsymbol{\zeta})^a} \frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial \zeta_k} \\ & + be^{\mathbf{x}_i^\top \boldsymbol{\beta}} \frac{aG(y_i|\boldsymbol{\zeta})^{a-1}}{1 - G(y_i|\boldsymbol{\zeta})^a} \frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial \zeta_k}, \quad k = 1, \ldots, q. \end{aligned}$$

$$\frac{\partial l(\boldsymbol{\theta}|D)}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i x_{ij} + b\log\{1 - G(y_i|\boldsymbol{\zeta})^a\}e^{\mathbf{x}_i^\top \boldsymbol{\beta}} x_{ij}, \quad \text{where} \;\; j = 1, \ldots, p.$$

The Kum-G-PO model:

Using the Equation (3.6), the general log-likelihood function of the Kum-G-PO models can be express

$$
\begin{aligned}
l(\boldsymbol{\theta}|D) = \sum_{i=1}^{n} & \delta_i[-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\} + \log(a) + \log(b) + \log[g(y_i|\boldsymbol{\zeta})] + (a-1)\log[G(y_i|\boldsymbol{\zeta})] \\
& + \log\{1-G(y_i|\boldsymbol{\zeta})^a\} + \mathbf{x}_i^\top\boldsymbol{\beta} - \log\left[1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right] \\
& + \log\left[1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right]^{-1} + 2(\kappa\log(\kappa) - \log(\Gamma(\kappa)) \\
& + (\kappa-1)(\log(b)+\log(a)) - \kappa(b+a).
\end{aligned}
$$

The elements of the score vector of the Kum-G-PO models is given by

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta}|D)}{\partial a} = & \sum_{i=1}^{n} \frac{\delta_i}{a} + \frac{\delta_i b G(y_i|\boldsymbol{\zeta})^a \log[G(y_i|\boldsymbol{\zeta})]}{1-G(y_i|\boldsymbol{\zeta})^a} + \delta_i\log[G(y_i|\boldsymbol{\zeta})] - \frac{G(y_i|\boldsymbol{\zeta})^a\log[G(y_i|\boldsymbol{\zeta})]}{1-G(y_i|\boldsymbol{\zeta})^a} \\
& + \frac{\delta_i e^{\mathbf{x}_i^\top\boldsymbol{\beta}} \exp\left[-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\right] bG(y_i|\boldsymbol{\zeta})^a \log[G(y_i|\boldsymbol{\zeta})]}{1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\left[1-G(y_i|\boldsymbol{\zeta})^a\right]} \\
& - \frac{\exp\left[-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\right] b\log[G(y_i|\boldsymbol{\zeta})]}{1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\left[1-G(y_i|\boldsymbol{\zeta})^a\right] e^{2\mathbf{x}_i^\top\boldsymbol{\beta}}} \\
& \times \frac{1}{\left[1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right)\right]^2} + \frac{\kappa-1}{a} - \kappa.
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta}|D)}{\partial \zeta_k} = & \sum_{i=1}^{n} \delta_i b \frac{aG(y_i|\boldsymbol{\zeta})^{a-1}}{1-G(y_i|\boldsymbol{\zeta})^a}\frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial\zeta_k} + \delta_i \frac{1}{g(y_i|\boldsymbol{\zeta})}\frac{\partial g(y_i|\boldsymbol{\zeta})}{\partial\zeta_k} + \delta_i(a-1)\frac{1}{G(y_i|\boldsymbol{\zeta})}\frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial\zeta_k} \\
& + \delta_i \frac{1}{1-G(y_i|\boldsymbol{\zeta})^a} aG(y_i|\boldsymbol{\zeta})^{a-1}\frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial\zeta_k} - \delta_i \frac{e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\}}{1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}} \\
& \times \frac{baG(y_i|\boldsymbol{\zeta})^{a-1}}{1-G(y_i|\boldsymbol{\zeta})^a}\frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial\zeta_k} + \frac{e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} baG(y_i|\boldsymbol{\zeta})^{a-1}}{\left(1-G(y_i|\boldsymbol{\zeta})^a\right)\left(1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right)} \\
& \times \frac{1}{\log\left[1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right]}\frac{\partial G(y_i|\boldsymbol{\zeta})}{\partial\zeta_k}.
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta}|D)}{\partial \beta_j} = & \sum_{i=1}^{n} \delta_i x_{ij} - \frac{\delta_i x_{ij}\left[\exp\{-b\log(1-G(y_i|\boldsymbol{\zeta})^a)\} - 1\right]}{1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}} \\
& - \frac{x_{ij}}{\log\left[1 + \left(\exp\{-b\log\{1-G(y_i|\boldsymbol{\zeta})^a\}\} - 1\right) e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right]^2}.
\end{aligned}
$$

The Kum-G-AFT model:

The log-likelihood function of the Kum-G-AFT models can be express using the Equation (3.7). So we can written by

$$\begin{aligned}
l(\boldsymbol{\theta}|D) = \sum_{i=1}^{n} &\delta_i \log(a) + \log(b) + \delta_i \log\left[g\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right] + \delta_i (a-1) \log\left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right] + \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} \\
&- \delta_i \log\left(1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a\right) + b \log\left[1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a\right] e^{\mathbf{x}_i^\top \boldsymbol{\beta}} + 2(\kappa \log(\kappa) - \log(\Gamma(\kappa)) \\
&+ (\kappa - 1)(\log(b) + \log(a)) - \kappa(b + a).
\end{aligned}$$

The elements of the score vector of the Kum-G-AFT models can be express by

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\theta}|D)}{\partial a} = \sum_{i=1}^{n} &\frac{\delta_i}{a} + \delta_i \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right] + \frac{\delta_i G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a \log\left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]}{1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a} \\
&- \frac{b e^{\mathbf{x}_i^\top \boldsymbol{\beta}} G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a \log\left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]}{1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a} + \frac{\kappa - 1}{a} - \kappa.
\end{aligned}$$

$$\frac{\partial l(\boldsymbol{\theta}|D)}{\partial b} = \sum_{i=1}^{n} \frac{\delta_i}{b} + b \log\left[1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a\right] e^{\mathbf{x}_i^\top \boldsymbol{\beta}} + \frac{\kappa - 1}{b} - \kappa.$$

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\theta}|D)}{\partial \zeta_k} = \sum_{i=1}^{n} &\delta_i \frac{1}{\left[g\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]} \frac{\partial \left[g\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]}{\partial \zeta_k} + \delta_i (a-1) \frac{1}{\left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]} \frac{\partial \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]}{\partial \zeta_k} \\
&+ \delta_i \frac{a \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]^{a-1}}{1 - \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]^a} \frac{\partial \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]}{\partial \zeta_k} + b e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \frac{a \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]^{a-1}}{1 - \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]^a} \frac{\partial \left[G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\right]}{\partial \zeta_k}.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\theta}|D)}{\partial \beta_j} = \sum_{i=1}^{n} &- \frac{\delta_i g'\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\left(\frac{y_i x_{ij}}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right)}{g\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)} - \frac{(a-1)\delta_i G'\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)\left(\frac{y_i x_{ij}}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right)}{G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)} \\
&+ \frac{\delta_i a G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a-1}\left(\frac{y_i x_{ij}}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right)}{1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a} - \frac{\delta_i (b-1) a G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^{a-1}\left(\frac{y_i x_{ij}}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}\right)}{1 - G\left(\frac{y_i}{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}|\boldsymbol{\zeta}\right)^a}.
\end{aligned}$$

Appendix B: The Bernstein Polynomial (BP) Model

The Bernstein Polynomials (BP) were originally proposed by Bernstein (1913) as a proof for the Weierstrass Approximation Theorem (WAT) in the unit interval Lorentz (1986). Following p. 148 of Bartle and Sherbert (2011), the WAT states [Weierstrass Approximation Theorem] Let $I = [a, b]$ and let $v : I \mapsto \mathbb{R}$ continuous over I. If $\varepsilon > 0$ is given, then there exists a polynomial function p_ε such that $|v(x) - p_\varepsilon(x)| < \varepsilon$ for all $x \in I$.

To understand the BP approximation, first consider an event A such as $\mathbb{P}(A) = x$, where $\mathbb{P}$ is a probability measure. Then, suppose that an experiment with m trials will be performed in such a way that, if the event A occurs k times, $0 \leqslant k \leqslant m$, a monetary amount equal to $v(k/m)$ will be paid to a hypothetical gambler. Thereby, a random variable K defined as the number of successes (the event A has happened) in m trials has a binomial distribution: $K \sim Bin(x, m)$, where $x \in [0, 1]$. Therefore, the probability of k occurrences for the event A and the expected value for a random variable $Q = v(K/m)$ representing the amount received by the gambler are given respectively by

$$\mathbb{P}(K = k) = \binom{m}{k} x^k (1-x)^{m-k},$$

$$\mathbb{E}_m(Q) = B_{(m)}(x) = \sum_{k=0}^{m} v\left(\frac{k}{m}\right)\binom{m}{k} x^k (1-x)^{m-k}, \quad x \in [0, 1].$$

From the relations in (5) and (5), combined with the Theorem 5, Bernstein proved that, given $\varepsilon > 0$, $|v(x) - \mathbb{E}_m(Q)| < \varepsilon$. In other words

$$v(x) = \lim_{m\to\infty} \mathbb{E}_m(Q) = \lim_{m\to\infty} \sum_{k=0}^{m} v\left(\frac{k}{m}\right)\binom{m}{k} x^k (1-x)^{m-k} = \lim_{m\to\infty} B_{(m)}(x).$$

Thus, the Bernstein Polynomial of degree m that approximates $v(x)$ is given by $B_m(x)$, where

$$b_{(k,m)}(x) = \binom{m}{k} x^k (1-x)^{m-k}$$

is the Bernstein basis. Note that each basis can be seen as a weight since, given the degree m, $b_{(k,m)}(x) \in (0, 1)$ for all k and

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

To accommodate functions restricted to any compact interval $[a, b]$, $a < b \in \mathbb{R}$, the result in (5) can be extended as (Farouki and Rajan (1987), p. 191)

$$B_{(m)}(x) = \sum_{k=0}^{m} v\left[a + \frac{k}{m}(b-a)\right] b_{(k,m)}\left(\frac{x-a}{b-a}\right), \quad x \in [a, b].$$

According to Carnicer and Peña (1993), the BP approximation has optimal shape-preserving property when compared to other polynomial approximations. The Section 5 of Farouki (2012) review paper lists many properties and algorithms associated with the Bernstein bases (a total of 18 topics), but four are of major concern for the construction of a BP survival model:

1. **Symmetry**: $b_{(k,m-k)}(x) = b_{(k,m)}(1-x)$;
2. **Recursion**: $b_{(k,m+1)}(x) = x b_{(k-1,m)}(x) + (1-x) b_{(k-1,m)}(x)$;
3. **Non-negativity**: $b_{(k,m)}(x) \geqslant 0$, $\forall\, x \in [0, 1]$, if $0 \leqslant k \leqslant m$;
4. **Basis Derivative**: $\dfrac{d}{dx} b_{(k,m)}(x) = m\left[b_{(k-1,m-1)}(x) - b_{(k-1,m)}(x)\right]$.

Following Panaro (2020), the properties above allow the construction of an approximation for the derivative of $B_{(k,m)}$ in (5) with respect to x, which provides

$$\begin{aligned}
\frac{d}{dx} B_{(m)}(x; v) &= \sum_{k=0}^{m} v\left(\frac{k}{m}\right)\binom{m}{k}\left\{k x^{k-1}(1-x)^{m-k} - (m-k)x^k(1-x)^{m-k-1}\right\} \\
&= m \sum_{k=0}^{m} v\left(\frac{k}{m}\right)\left[\binom{m-1}{k-1} x^{k-1}(1-x)^{m-k} - \binom{m-1}{k} x^k (1-x)^{m-k-1}\right] \\
&= m \sum_{k=0}^{m} v\left(\frac{k}{m}\right) b_{(k-1,m-1)}(x) - m \sum_{k=0}^{m} v\left(\frac{k}{m}\right) b_{(k,m-1)}(x) \\
&= m \sum_{i=-1}^{m-1} v\left(\frac{i+1}{m}\right) b_{(i,m-1)}(x) - m \sum_{k=0}^{m} v\left(\frac{k}{m}\right) b_{(k,m-1)}(x), \qquad \text{(B.1)}
\end{aligned}$$

where $i = k - 1$. By definition (Farouki, 2012), consider $b_{-1,m-1}(x) = b_{m,m-1}(x) = 0$. Then, (B.1) can be rewritten as

$$\frac{d}{dx} B_{(m)}(x; v) = m \sum_{i=0}^{m-1}\left\{v\left(\frac{i+1}{m}\right) - v\left(\frac{i}{m}\right)\right\} b_{(i,m-1)}(x) = m \sum_{i=0}^{m-1} \Delta v_i^{(1)} b_{(i,m-1)}(x),$$

where $\Delta v_i^{(1)} = v[(i+1)/m] - v[i/m]$ is the first-order difference of $v(x)$ at $x = i/m$.

Chang et al. (2005) noted that the finite BP approximation could be used to estimate both hazard and cumulative hazard functions of a survival model, since this last is positive and bounded. Assuming $t \in [0, \tau]$, where $\tau = \inf\{t : S(t) = 0\} < \infty$, let $H(t)$ be the target function for the BP approximation. Thereby, rewriting (B.1) with $a = 0$ and $b = \tau$, the BP approximation for the cumulative hazard function is expressed as

$$B_{(m)}(t; H) = \sum_{k=0}^{m} H\left(\frac{k}{m}\tau\right) b_{(k,m)}\left(\frac{t}{\tau}\right), \quad t \in [0, \tau],$$

and its first derivative with respect to the time t (approximating the hazard function), using (B.1), as

$$\begin{aligned}
\frac{d}{dt}B_{(m)}(t; H) &= \frac{m}{\tau}\sum_{i=0}^{m-1}\left\{H\left(\frac{i+1}{m}\tau\right) - H\left(\frac{i}{m}\tau\right)\right\} b_{(i,m-1)}\left(\frac{t}{\tau}\right) \\
&= \frac{m}{\tau}\sum_{k=1}^{m}\left\{H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right)\right\}\binom{m-1}{k-1} b_{(k-1,m-1)}\left(\frac{t}{\tau}\right) \\
&= \frac{m}{\tau}\sum_{k=1}^{m}\left\{H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right)\right\}\binom{m-1}{k-1}\left(\frac{t}{\tau}\right)^{k-1}\left(1-\frac{t}{\tau}\right)^{(m-1)-(k-1)} \\
&= \frac{1}{\tau}\sum_{k=1}^{m}\left\{H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right)\right\}\frac{\Gamma(m+1)}{\Gamma(m-k+1)\Gamma(k)}\left(\frac{t}{\tau}\right)^{k-1}\left(1-\frac{t}{\tau}\right)^{m-k} \\
&= \sum_{k=1}^{m}\left\{H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right)\right\}\left(\frac{1}{\tau}\right) f_B\left(\frac{t}{\tau}; k, m-k+1\right), \qquad \text{(B.2)}
\end{aligned}$$

where B denotes the Beta distribution with parameters $\alpha = k$ and $\beta = m - k + 1$. For simplicity, the cumulative hazards differences between braces and the Bernstein bases in (B.2) will be rewritten, respectively, as

$$\gamma_k = \left\{H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right)\right\}, \quad g_{(k,m)}(t) = \left(\frac{1}{\tau}\right) f_B\left(\frac{t}{\tau}; k, m-k+1\right).$$

Note that $\gamma_k > 0$, $k \in \{1, \ldots, m\}$, since $H(\cdot)$ is monotone increasing. As the coefficients γ_k do not depend on t, no information is given on the true cumulative hazard function and they should be estimated, compounding a vector $\boldsymbol{\kappa} = \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)'$ of BP parameters. Given a time t, its Bernstein bases can also be defined on a vector

$\boldsymbol{g}_m(t) = (g_{(1,m)}(t), \ldots, g_{(m,m)}(t))'$ of fixed non-negative quantities. Then, the hazard and cumulative functions are modeled as (Osman and Ghosh (2012), p. 561)

$$\begin{aligned} h(t|\boldsymbol{\gamma}) &= \boldsymbol{\gamma}'\boldsymbol{g}_m(t), \\ H(t|\boldsymbol{\gamma}) &= \int_0^t h(u, \boldsymbol{\gamma}) du = \boldsymbol{\gamma}'\boldsymbol{G}_m(t), \end{aligned}$$

where $\boldsymbol{G}_m(t) = (G_{(1,m)}(t), \ldots, G_{(m,m)}(t))'$, with

$$G_{(k,m)}(t) = \int_0^t g_{(k,m)}(u) du = \int_0^t f_B\left(\frac{u}{\tau}; k, m-k+1\right) d\left(\frac{u}{\tau}\right) \geqslant 0, \quad k \in \{1, \ldots, m\}.$$

Appendix C: The proportional hazards assumption in applications I, II, and III

In this section, we present the standardized Schoenfeld residuals along with the test of proportional hazards assumption of the Cox model for each covariate included in the model. This residuals was calculated using the `survminer::ggcoxzph` package. Thus, based on Figure C.2, we can see that all covariates do not reject the null hypothesis of the test of the proportional hazards. Thus, for the laryngeal cancer data set, the proportional hazards models can be used to analyze the effect of the covariates age and stage of the disease on the survival time of patients until death.

In Figure Figure C.3, it can be seen that only the covariate ph.karno presented evidence to reject the assumption of proportional hazards. Therefore, the proportional hazards models are not suitable for this data set, that is, it is expected that these models present inadequate adjustments to the lung cancer data set.

Finally, based on Figure C.4, it can be seen that there is evidence of rejection of the assumption of proportional hazards for all covariates considered in the veteran data set, in other words, the proportional hazards models present an inadequate fit for this veteran data set.

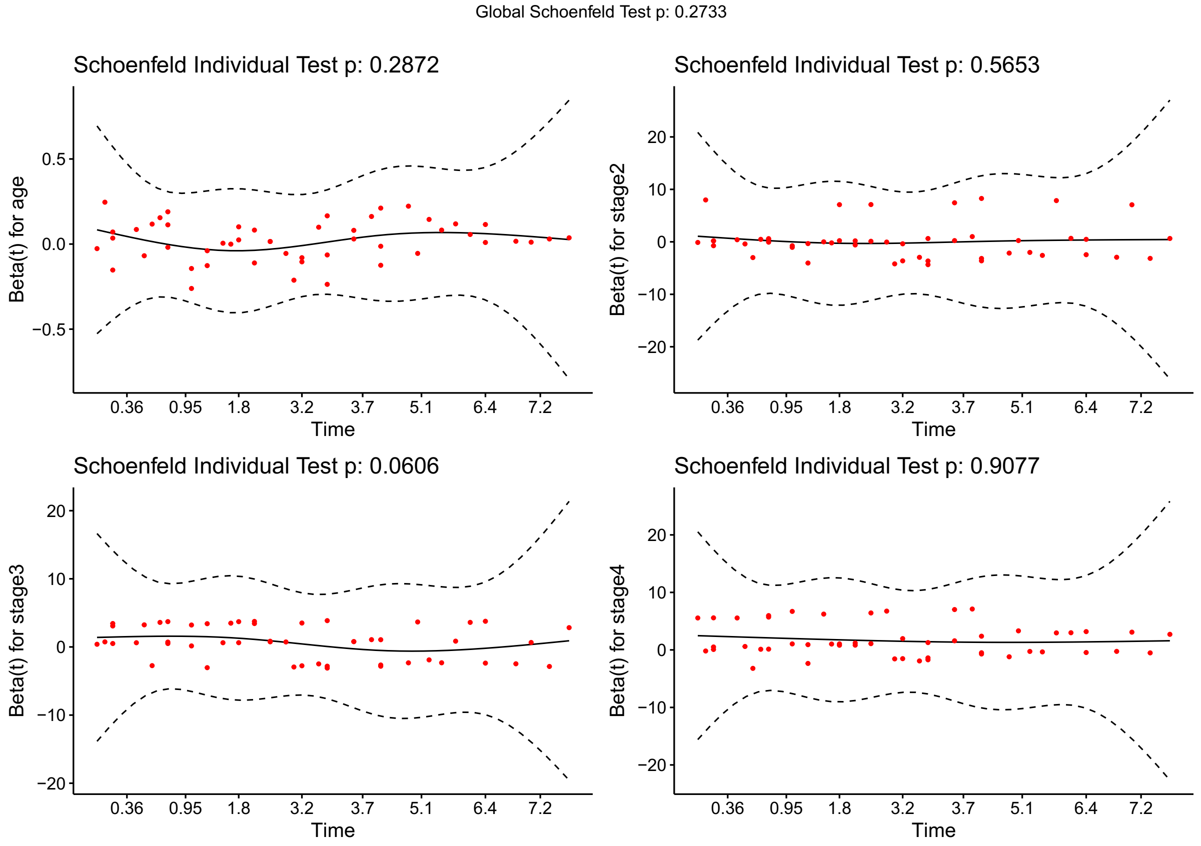


Figure C.2: The Standardized Schoenfeld residuals from the application I - larynx data set considering the test p-value for each covariate.

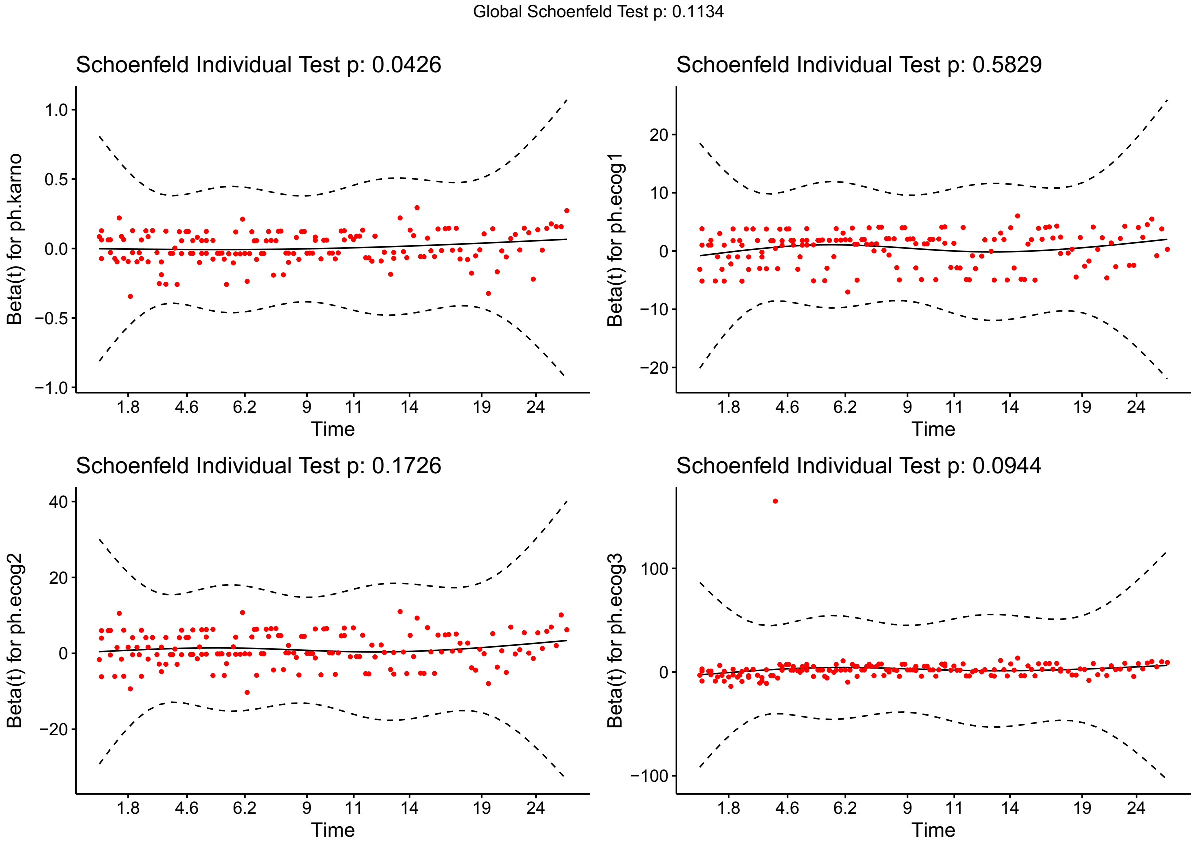


Figure C.3: The Standardized Schoenfeld residuals from the application II - lung data set considering the test p-value for each covariate.

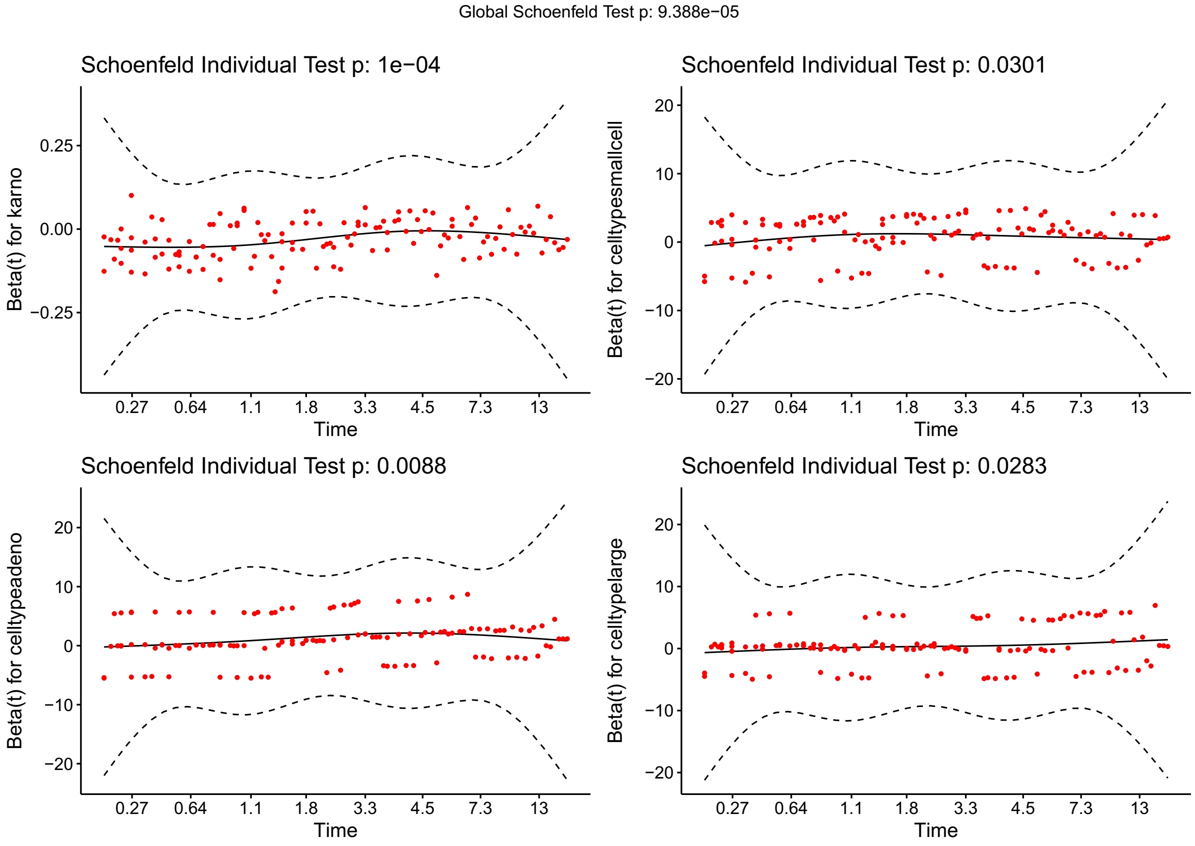


Figure C.4: The Standardized Schoenfeld residuals from the application III - veteran data set considering the test p-value for each covariate.

Appendix D: KGsurv package for survival data

In this section, we introduce the KGsurv package. The goal is to present an R package that has an easy routine and has all the support to avoid mistakes when using it. In this sense, this section was created to show the results presented by the KGsurv package. It is important to note that this package was created using Stan Team (2018). This package has several advantages of computational implementation, for example, obtain the results with fast computational speed, this speed occurs because the rstan package uses the C language. For these reasons, KGsurv package was built to have less intensive computational use so that users can get results with computers with few cores.

To install the KGsurv package, access https://github.com/CaioBalieiro/KGsurv and the following commands in R:

```
install.packages("devtools")
devtools::install_github("CaioBalieiro/KGsurv")
```

C.1 - Description of the kgreg function

```
kgreg(formula, data = NULL, hessian = TRUE, distG = c("exponential", "weibull", "gamma", "loglogistic", "lognormal"), regFamily = c("ph", "po", "aft"), penalty = 1, init = "random", ...)
```

Mandatory arguments:

- **formula**: an object of class ”formula” (or one that can be coerced to that class): a symbolic description of the model to be fitted, for example.

  ```
  formula = Surv(time, status) ~ gender
  ```

- **data**: an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which kgreg is called.

- **hessian**: logical; If TRUE (default), the hessian matrix is returned.

- `distG`: the G distribution used to derive the KG distribution, for instance: ”exponential”, weibull”, ”gamma”, ”loglogistic”, ”lognormal”.

- `regFamily`: the three regression models ”ph”, ”po”, ”aft”.

- `init`: initial values specification; default value is random.

- `penalty`: non-negative value passed to the penalty function; default value is 1.

- `...` : arguments passed to other methods.

The package was created to generalize its use, that is, any user can include his data sets, choose which model and distribution he/she wants to use. In this sense, the package was created using the `rstantools` package, that is, using the results of the Stan software, so it can be used in the R language in version $\geq$ 3.4.0. It is important to mention that it can be installed on Ubuntu, Mac, and Windows 64 bits. In the blocks below, some results will be presented using the `KGsurv::kgreg` function, considering its formulation and its respective output.

```
library(tidyverse)
library(KGsurv)

data(larynx, package = "KMsurv")
glimpse(larynx)

larynx <- larynx %>%
  mutate(
    stage = as.factor(stage),
    age = as.numeric(scale(age))
  )


glimpse(larynx)
```

```
mle <- kgreg(Surv(time, delta) ~ age + stage,
             data = larynx,
             distG = "loglogistic",
             regFamily = "ph",
             init = 0, penalty = 1)

summary(mle)

Survival regression model with Kumaraswamy-G baseline
   distributions
Call:
kgreg(formula = Surv(time, delta) ~ age + stage, data = larynx,
    regFamily = "ph", distG = "loglogistic", penalty = 1, init =
       0)

  n = 90  number of events =

         coef exp(coef) se(coef)      z Pr(>|z|)
age    0.2155    1.2405   0.1553 1.3876   0.1653
stage2 0.2108    1.2347   0.4621 0.4562   0.6482
stage3 0.6866    1.9869   0.3557 1.9303   0.0536 .
stage4 1.8525    6.3757   0.4241 4.3681   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

loglik = -141.4619   AIC = 298.9237  BIC = 318.9222
```

```
mle <- kgreg(Surv(time, delta) ~ age + stage,
             data = larynx,
             distG = "loglogistic",
             regFamily = "po",
             init = 0, penalty = 1)
summary(mle)
Survival regression model with Kumaraswamy-G baseline
   distributions
Call:
kgreg(formula = Surv(time, delta) ~ age + stage, data = larynx,
    regFamily = "po", distG = "loglogistic", penalty = 1, init =
       0)

  n = 90  number of events =

          coef  exp(coef)  se(coef)      z  Pr(>|z|)
age     0.2233     1.2502    0.2095 1.0659    0.2865
stage2  0.2684     1.3079    0.5906 0.4545    0.6495
stage3  1.2236     3.3994    0.5062 2.4172    0.0156 *
stage4  2.5701    13.0671    0.6242 4.1174    <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1          1

loglik = -141.0224   AIC = 298.0448  BIC = 318.0433
```

```
mle <- kgreg(Surv(time, delta) ~ age + stage,
             data = larynx,
             distG = "loglogistic",
             regFamily = "aft",
             init = 0, penalty = 1)

summary(mle)

Survival regression model with Kumaraswamy-G baseline
   distributions
```

```
Call:
kgreg(formula = Surv(time, delta) ~ age + stage, data = larynx,
    regFamily = "aft", distG = "loglogistic", penalty = 1, init =
        0)

  n = 90  number of events =

          coef  exp(coef) se(coef)       z  Pr(>|z|)
age     -0.1317    0.8766   0.1755 -0.7504    0.4530
stage2  -0.0942    0.9101   0.3940 -0.2391    0.8110
stage3  -0.8177    0.4414   0.5413 -1.5106    0.1309
stage4  -1.7749    0.1695   0.4843 -3.6649    0.0002 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

loglik = -142.1073   AIC = 300.2145  BIC = 320.213
```

In the box green, the results are of three options using the `KGsurv::kgreg` function. In the first case, we show the results of the Kum-llogis-PH model, in the second case Kum-llogis-PO, and Kum-llogis-AFT considering the penalty equal 1 for the laryngeal data set discussed in Chapter 4. Regarding the results presented, we have that, `coef` corresponds to the estimation of the parameters of the models, `exp (coef)` denotes the (HR), `se(coef)` corresponds to the standard error of the parameters considered by the model, `z` represents the Wald statistic, `Pr`$(> |z|)$ denotes the p-value associated with the Wald test.

It is important to highlight that, the results can be used for other distributions, for instance, Exponential, Gamma, Weibull, and Log-normal. As described in Chapter 3, it is observed that in many data sets, some models belonging to this family do not converge, one solution is to use penalty > 0, that is, we include a penalty in the log-likelihood of the models presented in `KGsurv` package. Also, when we use init $= r$, where r is the real number we define a more accurate initial guess for the search for all parameter estimates, for instance, in data sets considered in this work, we use $r = 0$. Therefore, the choice of r value is a great alternative to avoid the error of the models (by default init equal ”random”).

In addition KGsurv package presents some functions that are widely used in models already implemented in the R package such as: AIC, BIC, coef, confint, vcov, and coxsnell. To exemplify the Kum-llogis-PH model using the larynx data set.

```
mle <- kgreg(Surv(time, delta) ~ age + stage,
             data = larynx,
             distG = "loglogistic",
             regFamily = "ph",
             init = 0, penalty = 1)
AIC(mle)
 298.9237
BIC(mle)
 318.9222
coef(mle)
      age    stage2    stage3    stage4
0.2154815 0.2107623 0.6865571 1.8524926
attr(,"class")
[1] "coef.kgreg"
confint(mle)
               2.5%     97.5%
beta[1] -0.08888047 0.5198436
beta[2] -0.69496779 1.1164923
beta[3] -0.01064710 1.3837613
beta[4]  1.02132559 2.6836596
attr(,"class")
[1] "confint.kgreg"
coxsnell(mle)
 [1] 0.02622532 0.02919584 0.04529199 0.04760660 0.04890565
```

```
    0.05493653 0.06412331 0.09068648 0.10437034 0.11702129
    0.13782344
[12] 0.13883381 0.14115558 0.14533852 0.14650799 0.15114282
   0.15603189 0.16413696 0.18858155 0.18874863 0.19611580
   0.19663893
[23] 0.20932093 0.21031305 0.21644710 0.21878110 0.21906490
   0.22081925 0.23758758 0.24805974 0.27079471 0.27089143
   0.27503449
[34] 0.27554660 0.28056817 0.32262494 0.33595413 0.34191765
   0.34624138 0.35415044 0.35584259 0.38032486 0.39752455
   0.41820052
[45] 0.41843713 0.47788379 0.48288190 0.48355561 0.48634863
   0.50493166 0.52888436 0.55427088 0.57179223 0.57198294
   0.57975199
[56] 0.60119906 0.61041763 0.61656659 0.62023584 0.63778120
   0.64005662 0.65957901 0.66187050 0.67611309 0.67671232
   0.68366063
[67] 0.69383479 0.69558895 0.70699595 0.72677756 0.76111015
   0.76858816 0.79245789 0.84860793 0.86564422 0.87981462
   0.89107103
[78] 1.00002836 1.06903051 1.09197987 1.09556733 1.15597870
   1.42711148 1.45278705 1.47370571 1.48184185 1.80129211
   1.87003944
[89] 1.88796100 2.49836275
vcov(mle)
              beta[1]      beta[2]      beta[3]      beta[4]
                 gamma[1]     gamma[2]     gamma[3]     gamma[4]
beta[1]   0.024114858  0.008780830  0.004056729 -0.005049290
   -0.001388730 -0.001546975  0.004010302  0.001698628
beta[2]   0.008780830  0.213550880  0.068418492  0.068474080
   -0.004975975 -0.060781916  0.007326488  0.008095626
beta[3]   0.004056729  0.068418492  0.126538832  0.066785400
```

```
   0.008917128 -0.058294769  0.004017449 -0.004450726
beta[4]   -0.005049290  0.068474080  0.066785400  0.179837558
   -0.037868314 -0.084807123 -0.005739606  0.047090388
gamma[1] -0.001388730 -0.004975975  0.008917128 -0.037868314
   0.455624817  0.273044356  0.060520516 -0.338603719
gamma[2] -0.001546975 -0.060781916 -0.058294769 -0.084807123
   0.273044356  0.696948517  0.395528088 -0.159796256
gamma[3]  0.004010302  0.007326488  0.004017449 -0.005739606
   0.060520516  0.395528088  0.286124461 -0.014963409
gamma[4]  0.001698628  0.008095626 -0.004450726  0.047090388
   -0.338603719 -0.159796256 -0.014963409  0.275072886
attr(,"class")
[1] "vcov.kgreg"
```

In the table above, some functions were presented that are widely used in regression models, with `AIC`(mle) being the result of the AIC statistics, `BIC`(mle) being the result of the BIC measurement, `coef`(mle) are the maximum likelihood estimates of the regression coefficients, `confint`(mle) presents a matrix with the 95% confidence interval estimated for the regression coefficients, `coxsnell`(mle) calculates the Cox-Snell residuals and presents a plot of the residuals, and `vcov`(mle) presents a matrix with the estimated variance and covariance for the regression coefficients.

C.2 - Description of the explore_fits function

The KGsurv package, present the KGsurv::explore_fits function. This function presents a table containing the measures of regression family, distribution, penalty, fit, AIC, MSE, Code, and plot to 2 until 120 models implemented considered the penalty values (1e-03, 1e-02, 1e-01, 0, 1e+00,1e+01, 1e+02, 1e+03), the five models of the failure time (Kum-Exp, Kum-W, Kum-GA, Kum-llogis, and Kum-lnorm) and three regression families (PH, PO, e AFT).

```
explore_fits(formula, data, distG, regFamily, penalty, plot = FALSE, init
= "random", ...)
```

Mandatory arguments:

- `distG`: the G distribution used to derive the KG distribution, for instance: ”exponential”, ”weibull”, ”gamma”, ”loglogistic”, ”lognormal”.
- `regFamily`: The three regression models ”ph”, ”po”, ”aft”.
- `formula`: an object of class ”formula” (or one that can be coerced to that class): a symbolic description of the model to be fitted, for example;

  `formula = Surv(time, status) ~ age`
- `data`: an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which kgreg is called.
- `plot`: logical (default = TRUE); if TRUE, than the plot of the Cox-Snell residuals is displayed.
- `init`: Initial values specification.
- `penalty`: non-negative value passed to the penalty function; default value is (1e-03, 1e-02, 1e-01, 0, 1e+00,1e+01, 1e+02, 1e+03).
- `...` :Arguments passed to other methods.

```
all <-  explore_fits(formula = Surv(time, delta) ~ age + stage,
                     data = larynx,
                     distG = c("exponential", "weibull", "gamma",
                         "lognormal","loglogistic"),
                     regFamily = c("ph", "po", "aft"),
                     init = 0)
all
# A tibble: 120 x 8
   regFamily distG         penalty fit        AIC    MSE  code plot
   <chr>     <chr>           <dbl> <list>   <dbl>  <dbl> <int> <list
      >
 1 ph        lognormal        1000 <kgreg>  270. 0.0254     0 <gg>
 2 po        loglogistic      1000 <kgreg>  272. 0.0212     0 <gg>
 3 ph        loglogistic      1000 <kgreg>  273. 0.0265     0 <gg>
 4 aft       loglogistic      1000 <kgreg>  277. 0.0255     0 <gg>
 5 ph        lognormal         100 <kgreg>  279. 0.0254     0 <gg>
 6 po        loglogistic       100 <kgreg>  281. 0.0212     0 <gg>
 7 ph        loglogistic       100 <kgreg>  282. 0.0265     0 <gg>
 8 aft       loglogistic       100 <kgreg>  286. 0.0246     0 <gg>
 9 ph        lognormal          10 <kgreg>  288. 0.0251     0 <gg>
10 po        loglogistic        10 <kgreg>  289. 0.0210     0 <gg>
#     with 110 more rows
```

The green boxes above show examples using the `KGsurv::explore_fit` functions for the larynx data set. In them we can see that the output object has the tibble class, this class belongs to the `tidyverse` package. This class is more interesting than the list of classes in R, as it allows plot objects to be saved. It is important to remember that the functions have the argument plot = True as a default, so when using these functions a graph is generated, if you do not want to visualize just use plot = False.

C.3 - Description of the best_fits function

The `KGsurv::best_fits` function presents a table containing AIC, MSE, Code for the best model fits using the Pareto solution set. Besides, the `KGsurv::best_fits` shows a graph using the Pareto set and presents a plot of the Cox-Snell residuals for the best models containing the penalty, AIC, and MSE values. In the Pareto solution set graph, all the fitted models are in the form of points on an AIC vs MSE graph. The models that have the points in bold are the models that have the smallest fit measures (AIC and MSE) for the data set of the study.

`best_fits(fits, plotPareto = TRUE, plotResiduals = TRUE)`

Mandatory arguments:

- `fits`: the tibble class; output of the explore_fits function.
- `plotPareto`: logical (default = TRUE); if TRUE, than the plot of the Pareto set solution is displayed.
- `plotResiduals`: logical (default = TRUE); if TRUE, than the plot of the Cox-Snell residuals is displayed

```
all <-  explore_fits(formula = Surv(time, delta) ~ age + stage,
                     data = larynx,
                     distG = c("exponential", "weibull", "gamma",
                         "lognormal","loglogistic"),
                     regFamily = c("ph", "po", "aft"),
                     init = 0)
best_fits(all)
$models
# A tibble: 5 x 10
     id model           regFamily distG       penalty fit
        AIC    MSE  code plot
  <int> <chr>           <chr>     <chr>         <dbl> <list>   <dbl
     >  <dbl> <int> <list>
```

```
1     1 lognormal-ph   ph        lognormal      1000 <kgreg>
   270. 0.0254     0 <gg>
2     2 loglogistic-po po        loglogistic    1000 <kgreg>
   272. 0.0212     0 <gg>
3     3 loglogistic-po po        loglogistic     100 <kgreg>
   281. 0.0212     0 <gg>
4     4 loglogistic-po po        loglogistic      10 <kgreg>
   289. 0.0210     0 <gg>
5     5 lognormal-po   po        lognormal         1 <kgreg>
   296. 0.0166     0 <gg>

$pareto

$coxsnell
TableGrob (3 x 2) "arrange": 5 grobs
  z     cells    name           grob
1 1 (1-1,1-1) arrange gtable[layout]
2 2 (1-1,2-2) arrange gtable[layout]
3 3 (2-2,1-1) arrange gtable[layout]
4 4 (2-2,2-2) arrange gtable[layout]
5 5 (3-3,1-1) arrange gtable[layout]

attr(,"class")
[1] "kgreg.bestfits"
```

In the green box above, an example of the `KGsurv::best_fit` function using the larynx data set was presented. In it, we can notice that just like the `KGsurv::best_fits` function, a table is presented in the tibble class. The models presented in this table are the models that presented the best fits considering the Pareto solution set. In this dissertation, we present a package in R using regression models PH, PO, and AFT using the Kumaraswamy-G family of distributions. The package was called `KGsurv`, this package presents an easy and efficient application for several real data sets. It is important to highlight that, the functions presented here are widely used in data science, the objective is to make these tools more accessible to users, as they make the statistical analysis more robust, summarizing the information of the models presented in the `KGsurv` package. We hope that this package presents an attractive solution for survival data considering data with the right-censored.

Bibliography

Adepoju, K. and Chukwu, O. (2015), "Maximum likelihood estimation of the Kumaraswamy exponential distribution with applications," *Journal of Modern Applied Statistical Methods*, 14, 18.

Akaike, H. (1998), "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*, pp. 199–213, Springer.

Assunção, G. H. O. (2018), "Regressão espacial quantílica para previsão da velocidade do vento," Master's thesis.

Bartle, R. G. and Sherbert, D. R. (2011), *Introduction to Real Analysis*, John Wiley & Sons, 4th edn.

Bennett, S. (1983), "Analysis of survival data by the proportional odds model," *Statistics in medicine*, 2, 273–277.

Bernstein, S. N. (1913), "Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités," *Communications of the Kahrkov Mathematical Society*, 13, 1–2.

Brostom, G. (2014), "eha: event history analysis. R package version 2.4-2," .

Carnicer, J. M. and Peña, J. M. (1993), "Shape preserving representations and optimality of the Bernstein basis," *Advances in Computational Mathematics*, 1, 173–196.

Chacko, M. and Mohan, R. (2017), "Estimation of parameters of Kumaraswamy-Exponential distribution under progressive type-II censoring," *Journal of Statistical Computation and Simulation*, 87, 1951–1963.

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, 447–466.

Collett, D. (2015), *Modelling survival data in medical research*, CRC press.

Colosimo, E. and Giolo, S. (2006), *Análise de sobrevivência aplicada*, ABE - Projeto Fisher, Edgard Blücher.

Cordeiro, G. M. and de Castro, M. (2011), "A new family of generalized distributions," *Journal of statistical computation and simulation*, 81, 883–898.

Cordeiro, G. M., Ortega, E. M., and Nadarajah, S. (2010), "The Kumaraswamy Weibull distribution with application to failure data," *Journal of the Franklin Institute*, 347, 1399–1429.

Cordeiro, G. M., Pescim, R. R., and Ortega, E. M. (2012), "The Kumaraswamy generalized half-normal distribution for skewed positive data," *Journal of Data Science*, 10, 195–224.

Cordeiro, G. M., Ortega, E. M., and Silva, G. O. (2014), "The Kumaraswamy modified Weibull distribution: theory and applications," *Journal of Statistical Computation and Simulation*, 84, 1387–1411.

Cox, D. R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 187–202.

Cox, D. R. and Hinkley, D. V. (1979), *Theoretical statistics*, CRC Press.

Cox, D. R. and Snell, E. J. (1968), "A general definition of residuals," *Journal of the Royal Statistical Society: Series B (Methodological)*, 30, 248–265.

De Pascoa, M. A., Ortega, E. M., and Cordeiro, G. M. (2011), "The Kumaraswamy generalized gamma distribution with application in survival analysis," *Statistical methodology*, 8, 411–433.

D'Andrea, A., Rocha, R., Tomazella, V., and Louzada, F. (2018), "Negative Binomial Kumaraswamy-G Cure Rate Regression Model," *Journal of Risk and Financial Management*, 11, 6.

Elbatal, I. (2013), "The Kumaraswamy exponentiated Pareto distribution," *Economic Quality Control*, 28, 1.

Eugene, N., Lee, C., and Famoye, F. (2002), "Beta-normal distribution and its applications," *Communications in Statistics-Theory and methods*, 31, 497–512.

Farouki, R. T. (2012), "The Bernstein polynomial basis: A centennial retrospective," *Computer Aided Geometric Design*, 29, 379–419.

Farouki, R. T. and Rajan, V. (1987), "On the numerical condition of polynomials in Bernstein form," *Computer Aided Geometric Design*, 4, 191–216.

Ghosh, I. (2014), "The Kumaraswamy-half-Cauchy distribution: properties and applications," *Journal of Statistical Theory and Applications*, 13, 122–134.

Harrell Jr, F. E. et al. (2016), "rms: Regression modeling strategies," *R package version*, 5.

Hoggart, C. and Griffin, J. E. (2001), "A Bayesian partition model for customer attrition," .

Hubeaux, S. and Rufibach, K. (2014), "SurvRegCensCov: Weibull Regression for a Right-Censored Endpoint with a Censored Covariate," .

Jackson, C. H. (2016), "flexsurv: a platform for parametric survival modeling in R," *Journal of statistical software*, 70.

Jones, M. (2009), "Kumaraswamy's distribution: A beta-type distribution with some tractability advantages," *Statistical Methodology*, 6, 70–81.

Klein, J. P. and Moeschberger, M. L. (2006), *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media.

Konak, A., Coit, D. W., and Smith, A. E. (2006), "Multi-objective optimization using genetic algorithms: A tutorial," *Reliability engineering & system safety*, 91, 992–1007.

Kumaraswamy, P. (1980), "A generalized probability density function for double-bounded random processes," *Journal of Hydrology*, 46, 79–88.

Lawless, J. F. (2011), *Statistical models and methods for lifetime data*, vol. 362, John Wiley & Sons.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E. (1994), "Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group." *Journal of Clinical Oncology*, 12, 601–607.

Lorentz, G. G. (1986), *Bernstein Polynomials*, American Mathematical Society.

Meeker, W. Q. and Escobar, L. A. (2014), *Statistical methods for reliability data*, John Wiley & Sons.

Nadarajah, S. and Rocha, R. (2016), "Newdistns: An R package for new families of distributions," *Journal of Statistical Software*, 69, 1–32.

Nelson, W. (1990), "Accelerated life testing: statistical models, data analysis and test plans," *Accelerated life testing: statistical models data analysis and test plants.*

Osman, M. and Ghosh, S. K. (2012), "Nonparametric regression models for right-censored data using Bernstein polynomials," *Computational Statistics & Data Analysis*, 56, 559–573.

Panaro, R. V. (2020), "spsurv: An R package for semi-parametric survival analysis," .

Paranaíba, P. F., Ortega, E. M., Cordeiro, G. M., and Pascoa, M. A. d. (2013), "The Kumaraswamy Burr XII distribution: theory and practice," *Journal of Statistical Computation and Simulation*, 83, 2117–2143.

Prentice, R. L. (1973), "Exponential survivals with censoring and explanatory variables," *Biometrika*, 60, 279–288.

Santana, T., Ortega, E., Cordeiro, G., and Silva, G. (2012), "The Kumaraswamy-Log-Logistic Distribution," *Journal of statistical theory and applications*, 11, 265–291.

Saulo, H., Leão, J., and Bourguignon, M. (2012), "The kumaraswamy birnbaum-saunders distribution," *Journal of Statistical Theory and Practice*, 6, 745–759.

Scheike, T. H. and Zhang, M.-J. (2011), "Analyzing competing risk data using the R timereg package," *Journal of statistical software*, 38.

Schoenfeld, D. (1982), "Partial residuals for the proportional hazards regression model," *Biometrika*, 69, 239–241.

Struthers, C. A. and Kalbfleisch, J. D. (1986), "Misspecified proportional hazard models," *Biometrika*, 73, 363–369.

Team, S. D. (2018), "RStan: The R interface to Stan. R package version 2.18. 3," *Online: http://mc-stan. org.*

Therneau, T. (2014), "A package for survival analysis in S. R package version 2.37-7," .

Wald, A. (1943), "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical society*, 54, 426–482.

Wang, W., Zhang, L.-L., Chen, J.-J., and Wang, J.-H. (2015), "Parameter estimation for coupled hydromechanical simulation of dynamic compaction based on pareto multiobjective optimization," *Shock and Vibration*, 2015.