Instituto de MATEMÁTICA E ESTATÍSTICA
UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de conclusão do Curso de Bacharelado em Estatística

Modelo de Regressão GAMLSS para Análise de Sobrevivência

Celso Menoti da Silva

Orientadora: Profa. Dra. Silvana Schneider

Outubro, 2022.

CIP - Catalogação na Publicação

da Silva, Celso Menoti
Modelo de Regressão GAMLSS para Análise de Sobrevivência / Celso Menoti da Silva. -- 2022.
49 f.
Orientador: Silvana Schneider.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática e Estatística, Curso de Estatística, Porto Alegre, BR-RS, 2022.

1. Regressão GAMLSS. 2. Análise de Sobrevivência. 3. Fração de Cura. 4. Câncer de Pele. I. Schneider, Silvana, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Modelo de Regressão GAMLSS para Análise e Sobrevivência

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Porto Alegre, 14 de Outubro de 2022.

Profa. Dra. Silvana Schneider - UFRGS

Orientadora

_____

Profa. Dra. Paula Andreghetto Bracco - UFRGS

Examinador

_____

RESUMO

A análise por regressão é uma ferramenta de modelagem estatística muito utilizada no tratamento de dados. Os modelos mais tradicionais como a regressão linear simples ou os modelos lineares generalizados exigem suposições quanto ao tipo de distribuição da variável resposta e não são indicados para modelar a relações não-lineares. Para superar essas limitações, surgiram os modelos de regressão GAMLSS (Generalized additive models for location, scale and shape), que permite modelar os parâmetros de locação (μ), escala (σ) e forma (ν e τ) em função de covariáveis. Os modelos de regressão GAMLSS possibilitam o ajuste de distribuições que não pertencem à Família Exponencial e relações não-lineares entre variável resposta e covariáveis. Esses modelos de regressão podem ser extendidos para análise de sobrevivência. A presente investigação apresenta as definições dos modelos de regressão GAMLSS e uma extensão para dados de sobrevivência com fração de cura, proposta por Ramires et al (2019) [22]. O trabalho apresenta uma aplicação dessa extensão, na análise de sobrevivência de pacientes com câncer de melanoma do estado de São Paulo. A aplicação faz uma estimativa da fração de cura e demais parâmetros.

Palavra Chave: Regressão GAMLSS, Fração de Cura, Câncer de Melanoma.

ABSTRACT

Regression analysis is a statistical modeling tool widely used in data processing. More traditional models such as simple linear regression or generalized linear models require assumptions about the type of distribution of the response variable and are not suitable for modeling non-linear relation. To overcome these limitations, the GAMLSS (Generalized additive models for location, scale and shape) regression models emerged, which allow modeling the parameters of location (μ), scale (σ) and shape (ν and τ) as a function of covariates. The GAMLSS regression models allow the fitting of distributions that do not belong to the Exponential Family and non-linear relation between the response variable and covariates. These regression models can be extended for survival analysis. The present investigation presents the definitions of the GAMLSS regression models and an extension for survival data with cured fraction, proposed by Ramires et al (2019) [22]. This work also shows an application through the survival analysis on a dataset of patients with melanoma cancer diagnosed in the state of São Paulo. In this application, an estimate of the cured fraction is made using the GAMLSS model extended to the survival analysis.

Keywords: GAMLSS Regression, Cured Fraction, Skin Cancer.

CONTEÚDO

Lista de Figuras **iv**

Lista de Tabelas **v**

1 Introdução **1**

2 Modelo de Regressão GAMLSS **4**
2.1 O Modelo GAMLSS 5
2.1.1 Estimação do Modelo 6
2.1.2 Regressão GAMLSS no **R** 8
2.1.3 Inferência 9
2.2 Preditor Linear 11
2.2.1 Componente Paramétrico 11
2.2.2 Termo aditivo 12
2.2.3 Suavização por Splines - *smoothing splines* 13
2.3 Análise de Diagnóstico 15
2.3.1 Função plot() 16
2.3.2 Função wp() 16

3 GAMLSS para análise de sobrevivência **18**
3.1 Modelos de fração de cura 19
3.2 GAMLSS para o Modelo Weibull de Taxa de Cura 21

4 Análise de sobrevivência para o estudo sobre câncer de melanoma **24**

5 Conclusões e Perspetivas **33**

6 Apêndice **35**
6.1 Apêndice 1: Implementação no R 35
6.1.1 Exemplo 1: Dados de Internação 35
6.1.2 Exemplo 2: Células CD4 37
6.2 Apêndice 2: Implementação do GAMLSS para sobrevivência 43
6.2.1 Exemplo 4 43
6.3 Apêndice 3: Estadiamento da Doença 46

Referências Bibliográficas **46**

LISTA DE FIGURAS

2.1 Exemplo de worm plot. 17

4.1 Evolução da função de sobrevivência ao longo do tempo (em anos) obtido pelo método de Kaplan-Meier. 26
4.2 Evolução da função de sobrevivência ao longo do tempo (em anos) obtido pelo método de Kaplan-Meier para as variáveis: Sexo(A), Estágio(B), Categoria de Atendimento (C) e Tratamento (D) 27
4.3 Ajuste não linear para o parâmetro de localização μ em relação a Idade. 29
4.4 Estimativa das Fração de Cura (ν) pelo regressão GAMLSS em cada categoria. 30
4.5 Função de Sobrevivência da População ajustada. 30
4.6 Função de Sobrevivência da População ajustada por variável e categoria. 31
4.7 Analise de Resíduos: (a) Worm Plot (b) Distribuição dos Resíduos Quantílicos (c) QQPlot dos resíduos quantílicos 32

6.1 Os termos ajustados para média no Modelo 4 38
6.2 Os termos ajustados para o desvio no Modelo 4 39
6.3 O worm plot do modelo 4 39
6.4 Comparação entre a curva ajustada e os dados do Exemplo 2 40
6.5 Comparação entre a curva ajustada e os dados do Exemplo 2. 41
6.6 Worm plot por segmento do ajuste pelo polinômio fracionário para a média e linear para a dispersão. 43
6.7 Worm plot por segmento do ajuste pelo polinômio fracionário para a média e linear para a dispersão. 44
6.8 Relação entre as variáveis idade, sexo e tratam com a variável tempo 45
6.9 Estadiamento Clínico 8ª EDIÇÃO AJCC – 2017 (Gomes et al (2017) [12]) . . . 46

LISTA DE TABELAS

4.1 Descrição das variáveis utilizadas na análise. 25
4.2 Variáveis utilizadas em cada parâmetro. 27
4.3 Estimação dos coeficientes de ajuste para os parâmetros μ, σ e ν. 28

6.1 Valores de AIC e BIC calculados dos Modelos propospos por Rigby e Stasinopoulos (2020), [30] para os dados do problema apresentado por Gande et al (1996) [11] 37
6.2 Valores de AIC e BIC calculados dos Modelos propostos por Rigby e Stasinopoulos (2020), [30] para os dados do problema apresentado por Wade e Ades (1994) [32]. A coluna Grau corresponde ao grau do polinômio ortogonal do ajuste. 40
6.3 Valores de AIC e BIC calculados dos Modelos com expoente fracionário. A coluna *n* corresponde ao número de potência do ajuste. 42
6.4 Comparação entre as metodologias de análise de sobrevivência no R. 45

1 INTRODUÇÃO

A construção de modelos matemáticos e estatísticos que relacionam variáveis é uma ferramenta poderosa na análise e tratamento de dados. O modelo de regressão linear simples, por exemplo, ajusta uma reta entre a variável resposta e uma covariável, porém a aceitação desse ajuste como modelo estatístico para estimar médias, irá depender se a variável reposta é normalmente distribuída.

Nelder et al (1972) [21] propuseram os Modelos Lineares Generalizados (Generalised Linear Model - GLM), nessa metodologia a variável resposta deve pertencer à uma distribuição da Família Exponencial. Essa metodologia possibilitou ampliação nas possibilidades de ajuste, pois a média pode ser estimada através de uma função de ligação com o termo linear $\mathbf{X}\boldsymbol{\beta}$. Com o surgimento de técnicas de suavização, Hastie e Tibshirani (1990) [15] foram os primeiro a apresentá-las em modelos lineares generalizados e eles e chamaram de Modelos Aditivos Generalizados (Generalised Additive Model-GAM).

As metodologias GLM e GAM são conhecidas como métodos de estimação do parâmetro de localização da distribuição de probabilidade da variável resposta, isto é a média. A modelagem do parâmetro de escala da distribuição de probabilidade da variável resposta, que está relacionado à dispersão, foi modelado inicialmente por Harvei (1976) [14] que modelou variância da distribuição normal em função de variáveis. Rigby e Stasinopoulos (1996) [23] introduziram funções de suavização para modelar média e variância, mas ainda necessitando de que a variável resposta seguisse uma distribuição de probabilidade da Família Exponencial. Posteriormente, Rigby e Stasinopoulos (2005) [24] apresentam a metodologia GAMLSS (Generalized additive models for location, scale and shape) que permite modelar qualquer parâmetro da distribuição de probabilidade da variável resposta em função de covariáveis, para qualquer tipo de distribuição de probabilidade. Essa metodologia ainda permite implementar funções de suavização nas covariáveis.

O desenvolvimento de novas metodologias de modelagem estatística decorre da necessidade de superar limitações das metodologias usuais, para dar respostas à problemas atuais. O aprimoramento da modelagem estatística para a área de saúde se torna cada vez mais necessário, principalmente ao tratar de dados sobre doenças como câncer, em que um dos objetivos é avaliar o tempo de sobrevida e a probabildiade de cura.

O câncer é uma doenças em que há proliferação descontrolada de células levando à formação de um tecido anormal. O diagnóstico do câncer, ou neoplasia, acontece quando as células deformadas crescem de forma desordenada e se espalham. O câncer de pele é uma classe dessa doença muito comum. Segundo Esteva et al (2017) [10], um em cada cinco estadunidenses será diagnosticado com câncer de pele durante a vida. No Brasil, os registros de câncer de pele em 2020 pelo Ministério da Saúde mostram uma taxa de incidência de 4,03casos/10000hab em homens e 3,94casos/10000hab em mulheres ([5]).

Para o Grupo Brasileiro de Melanoma [6], a pele é o maior órgão do corpo humano e corresponde a cerca de 16% do peso do corpo. A pele é uma importante barreira de proteção às agressões do meio externo e atua na síntese de vitamina D. De acordo com Esteva et al (2017) [10], câncer de pele começa na epiderme, parte mais superficial da pele. Para os autores, o melanoma é um câncer de pele maligno, que se desenvolve nos melanócitos, as células que produzem melanina, e é a causa da maioria dos óbitos causados pela doença. Tem grande facilidade para se espalhar para outros órgãos, formando metástases.

Segundo Esteva et al (2017)[10], o câncer de pele detectado precocemente apresenta alta taxa de cura. Na modelagem estatística essa informação pode ser traduzida como Fração de Cura. Uma forma de avaliar a Fração de Cura é extender o modelo de regressão GAMLSS para análise de sobrevivência. Ramirez et al (2019) [22] apresentaram uma extensão que permite usar GAMLSS e ajustar os parâmetros da distribuição de probabilidade do tempo até o evento de interesse em função de covariávies. Esse modelo de regressão permite ainda estimar a probabilidade de cura.

O presente trabalho tem por objetivo apresentar as definições do modelo de regressão GAMLSS e sua extensão para Análise de Sobrevivência. Além disso, foi considerado um estudo com pacientes de diagnosticados com câncer de pele do estado de São Paulo. Através da metodologia

proposta por Ramires et al (2019) [22], foi possível estimar os parâmetros da distribuição de probabilidade do tempo de sobrevivência, considerando a Fração de Cura como um parâmetro dessa distribuição.

Especificamente o trabalho tem como objetivos:

- Discorrer detalhadamente sobre regressão GAMLSS;
- Apresentar a extensão dos modelos de regressão GAMLSS para Análise de Sobrevivência com fração de cura, proposta por Ramires et al (2019) [22], e estimar a função de sobrevivência;
- Estimar a Fração de Cura de pacientes com diagnóstico de câncer de melanoma em função das covariáveis.

A investigação proposta neste trabalho, para alcançar os objetivos, está dividida da seguinte forma:

- no Capítulo 2, apresenta-se a construção do modelo de regressão GAMLSS. São apresentados conceitos gerais sobre a metodologia, a função de verossimilhança penalizada e os suavizadores do tipo spline. Nesse capítulo também é apresentado como é feita a análise dos resíduos para a regressão GAMLSS;
- no Capítulo 3, aborda-se as questões relevantes à Análise de Sobrevivência. Este capítulo apresenta uma extensão da regressão GAMLSS para Análise de Sobrevivência com fração de cura;
- no Capítulo 4, são apresentados os resultados da aplicação da extensão do modelo GAMLSS para Análise de Sobrevivência com fração de cura, o que permitiu avaliar a fração de cura em função das covariáveis. A aplicação é feita com dados de câncer de melanoma obsevados pela secretaria de saúde do estado de São Paulo e compilados pela FOSP (Fundação Oncocentro de São Paulo);
- no Capítulo 5, são apresentadas as considerações finais e as perspectivas futuras.

2 MODELO DE REGRESSÃO GAMLSS

Entre as técnicas de modelagem para regressão, os Modelos Lineares Generalizados (GLM - Generalized Linear Model) e os modelos aditivos generalizados (Generalized Aditive Model - GAM) são amplamente difundidos. Em ambos os modelos a variável resposta deve assumir uma distribuição da Família Exponencial, na qual a média da variável resposta é modelada em função das variáveis explicativas. Nesses modelos a variância da variável resposta depende de uma constante de dispersão e da média através da função de variância. Assim, nos modelos GLM e GAM, a variância, assimetria e curtose não são modelados explicitamente em termos de variáveis explicativas, mas implicitamente através de sua dependência da média da variável resposta (McCulloch 2001 [20] e Hastie e Tibshirani (1990), [15]).

Os modelos GLM e GAM são técnicas de modelagem de regressão flexíveis, porém não permitem ajustes fora da Família Exponencial. Rigby e Stasinopoulos (2005), [24], propuseram o modelo aditivo generalizado para locação, escala e forma (Generalized additive models for location, scale and shape - GAMLSS), para o qual não é necessário que a distribuição da variável resposta pertença à Família Exponencial. Além disso, todos os parâmetros da distribuição da variável resposta podem ser modelados em função das covariáveis.

Para uma variável resposta Y, com densidade de probabilidade $f(y|\boldsymbol{\theta})$, no qual $\boldsymbol{\theta} = \begin{bmatrix} \mu & \sigma & \nu & \tau \end{bmatrix}$ é o vetor de parâmetros da distribuição, em que μ é a média, σ é a desvio, ν é a assimetria e τ é a curtose. O parâmetro de localização representa o centro da distribuição e é frequentemente a média, podendo ser a mediana ou moda. O parâmetro de escala está relacionado à dispersão da distribuição, podendo ser o desvio padrão ou o coeficiente de variação. Os parâmetros ν e τ estão ligados à forma da distribuição (Rigby e Stasinopoulos (2005), [24]).

No caso da variável aleatória Y com média e variância definidas, ν indica o grau de assimetria da distribuição de probabilidade. Se Y tem distribuição normal em torno de μ, então $\nu = 0$. Em geral, se $\nu < 0$, então a variável é assimétrica à esquerda; se $\nu > 0$, assimétrica à direita.

Já o coeficiente de curtose τ mede a intesidade dos picos de sua distribuição de probabilidade. Para $Y \sim N(0,1)$, $\tau = 3$. Dessa forma, se $\tau - 3$ é maior que zero, é usual indicar que tem mais pico no centro do que $N(0,1)$. Caso contrário, é usual indicar que a distribuição é mais achatada que $N(0,1)$ (Magalhães (2005), [18]).

2.1 O MODELO GAMLSS

Considerando uma variável aleatória Y, com n observações independentes, com densidade de probabilidade $f(y|\boldsymbol{\theta})$, em que $\boldsymbol{\theta}$ é o vetor com p componentes. Segundo Rigby e Stasinopoulos (2005), [24] o modelo GAMLSS para o parâmetro θ_k, $1 \leq k \leq p$, da função de densidade de probabilidade da variável Y pode ser ajustado por $J_k = J_k^{'} + J_k^{''}$ variáveis explicativas, dado por

$$g_k(\boldsymbol{\theta}_k) = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k^{''}} \mathbf{h}_{jk}(x_{jk}), \tag{2.1}$$

sendo:

- $k \in \{1, 2, \cdots p\}$;
- $\mathbf{X}_k$ é a matriz de ordem $n \times J_k^{'}$;
- $\boldsymbol{\beta}_k$ é o vetor de parâmetros do ajuste de tamanho $J_k^{'}$;
- $\mathbf{X}\boldsymbol{\beta}_k$ é a parcela paramétrica do preditor linear;
- $\mathbf{h}_{jk}(x_{jk})$ são funções de suavização não-paramétricas das variáveis explicativas;
- $\boldsymbol{\eta}_k = g_k(\boldsymbol{\theta}_k)$ é o preditor linear, sendo um vetor de tamanho n;
- $g(.)$ é a função de ligação, assim $\theta_k = g^{-1}(\eta_k)$. No **R**, o padrão da biblioteca **gamlss** é atribuir uma função de ligação compatível ao espaço paramétrico. Mais especificamente:
  - Ligação Identidade: θ_k é qualquer valor real;
  - Ligação Logarítmica: $\theta_k > 0$;
  - Ligação Logito: $0 < \theta_k < 1$.

Os conjuntos de variáveis explicativas podem ser semelhantes ou diferentes para cada um dos parâmetros da distribuição, cuja as relações podem ser consideradas como funções lineares ou funções de suavização, ou ambas [22].

Se não forem implementadas suavizações no modelo de Equação (2.1), o modelo é classificado como GAMLSS paramétrico, caso contrário é classificado como GAMLSS semiparamétrico (Rigby e Stasinopoulos (2005), [24]). No caso paramétrico, a Equação (2.1) pode ser reescrita por

$$g_k(\boldsymbol{\theta}_k) = \mathbf{X}_k\boldsymbol{\beta}_k. \tag{2.2}$$

Na regressão GAMLSS definida pela Equação (2.1) é possível implementar funções de suavizações P-spline $h_j(x_j)$, que de acordo com Rigby e Stasinopoulos (2014), [26], apresentam as seguintes condições:

- Cada função de suavização é modelada como uma função spline de regressão, isto é , $h_j(x_j) = \mathbf{Z}_j\boldsymbol{\gamma}_j$, na qual $\mathbf{Z}_j$ é uma matriz na base B-spline para as variáveis exploratórias e $\boldsymbol{\gamma}_j$ é um vetor de parâmetros;

- Os parâmetros $\boldsymbol{\gamma}_{jt}$ são os coeficientes das funções de suavização e são estimados sujeitos a uma penalidade $\lambda_j \sum_k^q \left(\Delta^k\gamma_{jt}\right)^2$, em que Δ é um operador de diferença.

Conforme Rigby e Stasinopoulos (2007), [25] o vetor de coeficientes $\boldsymbol{\gamma}_j$ que possuem algumas restrições estocásticas impostas pelo fato de $\mathbf{D}\boldsymbol{\gamma}_j \sim N(0,\lambda^{-1}\mathbf{I})$, na qual $\mathbf{D}$ é uma matriz de tamanho $q \times r$ que corresponde a k-ésima diferença do vetor q-dimensional γ. No **R**, é possível implementar um P-spline em uma variável na fórmula da função *gamlss* com o comando *ps()*.

2.1.1 ESTIMAÇÃO DO MODELO

A estimação dos parâmetros é feita via Máxima Verossimilhança. Sendo assim, para uma amostra de tamanho n de uma variável Y, no qual todos y_i's, $i = 1, \cdots, n$, são independentes e identicamente distribuídos, tal que $Y \sim D(\mu, \sigma, \nu, \tau)$. Então, o logaritmo da função de verossimilhança para um modelo paramétrico é dado por

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left(f_Y(y_i|\beta)\right). \tag{2.3}$$

As funções h_{jk} são suavizações P-splines, que de acordo com Rigby e Stasinopoulos (2014), [26] são definidas por definidas por $\mathbf{h}(x) = \mathbf{Z}\boldsymbol{\gamma}$. Para parâmetros de suavização fixos λ_{kj}, os parâmetros $\boldsymbol{\beta}$ e os efeitos aleatórios $\boldsymbol{\gamma}$, são estimados via máxima verossimilhança penalizada, dada por

$$l_p = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_{jk}^{t}\mathbf{P}_{jk}\gamma_{jk}, \tag{2.4}$$

sendo $\mathbf{P}_{jk}$ uma matriz simétrica que pode depender de um vetor de parâmetros de suavização.

É possível usar dois algoritmos para maximizar a função de verossimilhança penalizada da Equação (2.4) em relação aos parâmetros, que são os algoritmos CG e RS. Para Rigby e Stasinopoulos (2020), [30], o algoritmo RS é preferível, pois o CG é instável, principalmente no início das iterações, e diverge facilmente. Segundo os autores, algoritmo RS é geralmente mais estável e mais rápido, por essa razão é definido como padrão no pacote *gamlss()* do **R**.

O algoritmo RS requer valores iniciais para os parâmetros θ´s, que podem ser mudados pelo usuário. A ideia é fazer ajustes considerando pesos, $\mathbf{w}_k$, e uma variável resposta modificada, até a convergência da deviance global (Rigby e Stasinopoulos (1996), [23]). A variável resposta modificada (iterada) para ajustar θ_k é dada por

$$\mathbf{z}_k = \boldsymbol{\eta}_k + \mathbf{w}_k^{-1} \circ \mathbf{u}_k \tag{2.5}$$

em que:

- $\mathbf{z}_k$, $\boldsymbol{\eta}_k$, $\mathbf{w}_k^{-1}$ e $\mathbf{u}_k$ são vetores de comprimento n;
- $\mathbf{w}_k^{-1} \circ \mathbf{u}_k$ é o produto Hadamard elemento a elemento;
- $\boldsymbol{\eta}_k$ é o vetor de preditores do k-ésimo vetor paramétrico;
- $u_k = \frac{\partial l}{\partial \eta_k}$ é a função escore.

Os $\mathbf{w}_k$ são os pesos iterativos definido como

$$\mathbf{w}_k = -f_k \circ \left(\frac{\partial \theta_k}{\eta_k}\right) \circ \left(\frac{\partial \theta_k}{\eta_k}\right). \tag{2.6}$$

Há três diferentes formas de definir f_k , dependendo da informação disponível para a distribuição específica:

- $E\left(\frac{\partial^2 l}{\partial \theta_l^2}\right)$ se a esperança existe, levando ao algoritmo *Scoring de Fisher*;
- $\frac{\partial^2 l}{\partial \theta_l^2}$ levando ao algoritmo *Scoring Newton-Raphson*;
- $-\left(\frac{\partial l}{\theta_k}\right) \circ \left(\frac{\partial l}{\theta_k}\right)$levando ao algoritmo *Scoring quasi Newton*.

O processo é repetido até que não haja mudança na deviance global. Para finalizar, o algoritmo RS realiza a estimação dos parâmetros β e γ pelo algoritmo *backfitting* que é uma versão do algoritmo de Gauss-Seidel (Rigby e Stasinopoulos (1996), [23]).

2.1.2 REGRESSÃO GAMLSS NO R

A linguagem de programação **R** utilizada em análise de dados, dispõe do pacote **gamlss** que contém a função *gamlss()*, que permite usar essa metodologia. A função *gamlss()* aceita as fórmulas do tipo *glm()* e funções de suavização.

Para a implementação da função *gamlss()*, a distribuição da variável resposta Y, $f(y|\boldsymbol{\theta})$, necessita que $\log[f(y|\boldsymbol{\theta})]$ e suas derivadas de primeira e segunda ordem em relação aos parâmetros sejam computáveis. As distribuições que podem ser usadas na função *gamlss()* estão parametrizadas considerando os parâmetros de locação, escala e forma, respetivamente, isto é $\boldsymbol{\theta} = \begin{bmatrix} \mu & \sigma & \nu & \tau \end{bmatrix}$.

No caso de distribuições com outros tipos de parâmetros é necessário a reparametrização. Por exemplo, a distribuição Gama pode ser escrita em função dos parâmetros α e β, dada por

$$f(x|(\alpha,\beta)) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}. \tag{2.7}$$

Nesse caso é necessário a reparametrização. Para a distribuição gama, a reparametrização é feita da seguinte forma: $\mu = \frac{\alpha}{\beta}$ e $\sigma^2 = \frac{\beta^2}{\alpha}$. Assim a Equação (2.7) pode ser reescrita por

$$f(x|(\mu,\sigma)) = \frac{x^{1/\sigma^2-1} e^{-x/(\mu\sigma^2)}}{(\mu\sigma^2)^{1/\sigma^2}\Gamma(1/\sigma^2)}. \tag{2.8}$$

Como no GAMLSS cada parâmetros tem um preditor linear, então é necessário uma função de ligação entre o preditor e o parâmetro. No **R** é possível consultar quais os parâmetros de uma distribuição e qual a função de ligação entre o parâmetro e o preditor, pelo comando *gamlss.family()*. Por exemplo, para a distribuição beta (BE), o *gamlss.family* = *BE* indica que a distribuição tem dois parâmetros, μ e σ, e que ambos têm o logit como função de ligação.

2.1.3 **INFERÊNCIA**

A inferência nos ajustes GAMLSS é baseada predominantemente na teoria da Máxima Verossimilhança. No caso paramétrico, os erros padrões, intervalos de confiança e testes de hipótese, são estritamente válidos assintoticamente. Nos casos em que o modelo contém termos com funções de suavização, os erros padrões são subestimados, os intervalos de confiança têm força inferior à nominal e os testes de hipótese têm o erro do tipo I inflacionado (Rigby et. al. (2020), [30]). O **R** apresenta uma nota de saída reforçando que o uso de suavizadores no modelo implica em cautela no tratamento do erro padrão. Se o preditor (linear) contiver apenas termos lineares (sem suavização), então o erro padrão do preditor (linear) é calculado como no GLM. Quando a função de ligação não é a identidade, então o erro padrão é obtido pelo método delta. Na presença de suavizadores, o parcote *gamlss* calcula o erro padrão usando a metodologia apresenta por Chambers e Hastei (1992), [4].

A inferência baseada na verosimilhança em GAMLSS resulta nas propriedades gerais de Estimadores de Máxima Verossimilhança (EMVs). O vetor de parâmetros $\boldsymbol{\theta} = \begin{bmatrix} \mu & \sigma & \nu & \tau \end{bmatrix}$ apresenta os respectivos coeficientes de regressão: $\boldsymbol{\beta} = \begin{bmatrix} \beta_\mu & \beta_\sigma & \beta_\nu & \beta_\tau \end{bmatrix}$. Dessa forma, $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, [I_F(\boldsymbol{\beta})]^{-1}\right)$, em que $I_F(\boldsymbol{\beta})$ é a matriz informação esperada de Fisher, definida por

$$I_F(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right). \tag{2.9}$$

Dentre os métodos para obter intervalos de confiança para os parâmetros de regressão em GAMLSS destacam-se: Intervalos do tipo Wald, Intervalos baseados na verossimilhança perfilada e Intervalos de confiança bootstrap.

Os intervalos de confiança do tipo Wald são construídos com base na distribuição assintótica

dos EMVs. Assim para um particular parâmetro β, o intervalo de confiança fica definido por

$$\widehat{\beta} \pm z_{\alpha/2} EP(\widehat{\beta}), \tag{2.10}$$

em que $z_{\alpha/2}$ é o quantil da distribuição normal padrão e $EP(\widehat{\beta})$ é o erro padrão dos parâmetros de regressão. Os erros padrões são usualmente obtidos tomando a raiz quadrada dos elementos da diagonal da matriz de covariâncias dos estimadores (inversa da matriz informação).

Intervalo de confiança baseado na verossimilhança perfilada para o parâmetro β_k é definido pelo conjunto dos valores de β_0 tais que $-2\left[L(\beta_0, \widehat{\Psi}(\beta_0)) - L(\beta_k, \widehat{\Psi})\right] < \chi^2_1(\alpha)$, sendo $L(\beta_0, \widehat{\Psi}(\beta_0)$ a verossimilhança maximizada para $\beta_k = \beta_0$, $L(\beta_k, \widehat{\Psi})$ a verossimilhança maximizada de forma irrestrita e Ψ o conjunto dos demais parâmetros do modelo.

Intervalos de confiança bootstrap são obtidos a partir da simulação de novas amostras, relacionadas, de alguma forma, à amostra original. As três principais modalidades de simulação bootstrap são as seguintes:

- As novas amostras são simuladas do modelo especificado, substituindo os parâmetros pelas respectivas estimativas de máxima verossimilhança;

- As novas amostras são selecionados com reposição da amostra original.

Predição, no contexto de GAMLSS, pode ser aplicada a qualquer um dos parâmetros da distribuição $\boldsymbol{\theta}$. Assim tendo que $g(\theta_k) = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots \widehat{\beta_m} x_m$, por exemplo, então

$$\theta_k = g^{-1}\left(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_m} x_m\right). \tag{2.11}$$

Um intervalo de confiança assintótico, para θ_k pode ser determinado calculando, inicialmente, um IC assintótico para $g(\theta_k)$, ou seja,

$$\widehat{\theta_k} \pm z_{\alpha/2} EP(\widehat{\theta_k}) \tag{2.12}$$

e posteriormente aplicando a função inversa g_k^{-1} aos limites obtidos.

2.2 PREDITOR LINEAR

O preditor linear $\boldsymbol{\eta}_k(.)$ de um parâmetro θ_k pode ser composto de por duas parcelas: a componente paramétrica $\mathbf{X}_k\boldsymbol{\beta}_k$ e o termo aditivo $\mathbf{h}(x)$.

2.2.1 COMPONENTE PARAMÉTRICO

Em GAMLSS a parcela $\mathbf{X}_k\boldsymbol{\beta}_k$ do preditor linear para um parâmetro θ_k corresponde a componente paramétrica. Por exemplo, se X corresponde a uma variável numérica a forma mais simples de inseri-la ao modelo é através de seu efeito linear, na forma $\eta = \beta_0 + \beta_1 x$. No caso de variável categórica f com m níveis, então é possível inserir ao preditor por $\eta = \beta_0 + \beta_1 I(f = a_2) + \cdots + \beta_{m-1} I(f = a_m)$, sendo que $I(f = a_k) = 1$. A incorporação de variáveis *dummy* deve respeitar as condições necessárias para evitar multicolinearidade. Nesse caso, para m fatores são criadas $m - 1$ variáveis indicadoras $I(.)$ (Rigby et. al. (2020), [30]).

O preditor pode acomodar simultaneamente diversas covariáveis numéricas e/ou fatores e efeitos multiplicativos podem ser utilizados como forma de incorporar efeitos de interação entre variáveis. Interações podem envolver duas ou mais variáveis numéricas, ou dois ou mais fatores, ou ainda variáveis numéricas e fatores. Além disso, podemos considerar como efeitos multiplicativos termos definidos como potências de variáveis.

Para Rigby et. al. (2020), [30], polinômios são a forma mais simples de modelar relações não-lineares em regressão. Um preditor baseado num polinômio de grau p para uma variável X tem a seguinte forma: $\eta = \beta_0 + \beta_1 x + \beta_2 x^2 \cdots + \beta_p x^p$. Nesse caso, as colunas da matriz do modelo ($\mathbf{X}$), correspondentes a um vetor de 1 's, $x, \cdots, x^p$, formam uma base polinomial.

Modelos baseados em polinômios podem apresentar alguns problemas. Os valores de x^p podem aumentar rapidamente (ou diminuir rapidamente) gerando problemas numéricos e de multicolinearidade. Uma forma de evitar os problemas mencionados é trocar a base polinomial padrão por uma base polinomial ortogonal. Polinômios ortogonais produzem valores ajustados idênticos aos polinômios não ortogonais (se ambos tiverem a mesma ordem) (Rigby et. al. (2020), [30]).

Outra forma de produzir melhores ajustes com menor quantidade de termos é usar polinômios fracionários, isto é, polinômios de potência fracionária. No R a função *fp()* é uma im-

plementação dos polinômios fracionários introduzidos por Royston and Altman (1994)[28]. É possível definir o número de termos de um polinômio fracionário pelo argumento **npoly**.

Por exemplo se "npoly = 3"o modelo fica: $\eta = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2} + \beta_3 x^{p_3}$. Nesse caso p_1, p_2, $p_3 \in -2, -1, -1/2, 0, 1/2, 1, 2, 3$. Se duas potências são iguais, por exemplo $p_1 = p_2 = c$, então $\beta_1 x^c$ e $\beta_2 x^c log(x)$ são incluídos ao modelo. Caso as três potências sejam iguais então $\beta_1 x^c$, $\beta_2 x^c log(x)$ e $\beta_3 x^c \, (log(x))^2$ são incluídos (Rigby et. al. (2020), [30]).

2.2.2 **TERMO ADITIVO**

O termo aditivo $\sum_{j=1}^{J_k} \mathbf{h}_{jk}(x_{jk})$ pode modelar uma variedade de termos, como suavização e termos de efeito aleatório, bem como termos que são úteis para análise de séries temporais. Diferentes termos aditivos que podem ser incluídos no GAMLSS.

Para ilustrar um termo aditivo com diversos elementos, Rigby e Stasinopoulos (2020), [30] apresenta um exemplo de um estudo com cinco variáveis exploratórias quantitativas, x_1, x_2, x_3, x_4 e x_5 e uma variável categórica f em três níveis. Segundo os autores o preditor linear pode ser

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 I(f = 2) + \beta_4 I(f = 3) + \beta_5 x_1 x_2 + \beta_6 x_1 I(f = 2) + \beta_7 x_1 I(f = 3)$$
$$+ s_1(x_3) + s_2(x_2) + \beta_8 x_4 s_3(x_1) + s_4(x_1) I(f = 2) + s_5(x_1) I(f = 3) + s_6(x_3, x_5),$$

em que:

- β_0 é o termo constante;
- $\beta_1 x_1 + \beta_2 x_2$ são termos aditivos lineares de variáveis quantitativas;
- $\beta_3 I(f = 2) + \beta_4 I(f = 3)$ é o efeito principal da variável qualitativa;
- $\beta_5 x_1 x_2$ é a interação entre as variáveis x_1 e x_2;
- $\beta_6 x_1 I(f = 2) + \beta_7 x_1 I(f = 3)$ é a interação linear entre a variável x_1 e a variável qualitativa;
- $s_1(x_3) + s_2(x_2)$ são termos aditivos de suavização para x_3 e x_2;
- $x_4 s_3(x_1)$ termo da interação da variável x_4 com a suavização de x_1;

- $s_4(x_1)I(f = 2) + s_5(x_1)I(f = 3)$ é uma interação da variável qualitativa com termos de suavização para x_1;

- $s_6(x_3, x_5)$ interação suave das variáveis x_3 e x_4.

Os suavizadores $s(.)$ podem ser univariados ou multivariados e caracterizam o termos aditivos na fórmula de modelo GAMLSS. Os suavizadores modelam os principais efeitos não lineares das variáveis explicativas em um parâmetro da distribuição da variável resposta, Y.

De acordo com Rigby e Stasinopoulos (2020), [30] os suavizadores penalizados são os suavizadores mais importantes dentro da família GAMLSS de suavizantes por causa de sua flexibilidade e o fato de que eles podem ser aplicados em uma variedade de diferentes situações. Todos os suavizadores considerados pelos autores são considerados como o solução para o problema de minimização de mínimos quadrados, onde certas restrições quadráticas são aplicadas aos parâmetros.

2.2.3 **SUAVIZAÇÃO POR SPLINES - *SMOOTHING SPLINES***

Com a finalidade de modelar efeitos não lineares o termo aditivo do preditor linear é composto por suavizadores não-paramétricos. Para Rigby e Stasinopoulos (2020), [30] a necessidade desses suavizadores ocorre porque ajuste de modelos polinomiais requerem a especificação da função de regressão global para todo o intervalo de valores de x. Dependendo da complexidade dos dados, nem sempre uma única especificação de polinômio é capaz de descrever a variação dos dados para todo valor de X. O ajuste de polinômios de elevada ordem, nesses casos, não é recomendado, devido ao elevado número de parâmetros, risco de *overfitting* e inflação da variância dos estimadores.

Para captar comportamentos não-lineares e aleatórios no modelo GAMLSS é possível incluir splines como termos aditivos. Uma função $s(x)$ é um spline de grau m definida em uma partição do intervalo $[a, b]$ baseada em K nós. Se $s(x)$ é um polinômio de grau m em cada subintervalo da partição, $s(x)$ tem $m-1$ derivadas sucessivas contínuas no infímo do subintervalo. Um spline de ordem m definido em uma partição do intervalo $[a,b]$ em K nós requer $K + m + 1$ parâmetros. Um modelo de regressão que usa spline polinomial de grau m como modelo aditivo

fica definido por

$$\eta = \mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^{K} \beta_j (x - b_j)^m I(x > b_j), \tag{2.13}$$

em que $I(.)$ é uma função indicadora.

A suavização por splines penalizados, ou P-spline, definidos por Eiler e Marx (1996) [9] são polinômios definidos por funções de base B-spline na variável explicativa. Dessa forma, é implementado um controle local da curva, de forma que a alteração de um ponto de controle modifica somente a região dos pontos vizinhos mais próximos em função da ordem de continuidade.

Uma curva B-spline, necessita de um vetor de nós, das funções de base e um conjunto de pontos de controle. De acordo com Eiler e Marx (1996) [9], uma função de base é denotada por $N_{ip}(x)$, que significa que é a i-ésima função de base B-spline de grau p (ordem $p + 1$). Uma curva B-spline de grau p pode ser definida em n pontos de controle por

$$C(x) = \sum_{i=1}^{n} N_{ip}(x) P_i, \tag{2.14}$$

sendo P_i o i-ésimo ponto de controle e N_{ip} a função de base de grau p definidas para um vetor de $m + 1$ nós. As funções de base são definidas da seguinte forma:

$$N_{i,0}(x) = \begin{cases} 1 & \text{se } x_i < x < x_{i+1} \\ 0 & \text{caso contrário} \end{cases}$$

$$N_{i,p}(x) = \frac{x - x_i}{x_{i+p} - x_i} N_{i,p-1}(x) + \frac{x_{i+p+1} - x}{x_{i+p+1} - x_{i+1}} N_{i+,p-1}(x).$$

Para Rigby e Stasinopoulos (2020), [30] qualquer função dada pela Equação (2.13) de um dado grau D e para um dado intervalo que contém x pode ser representado exclusivamente por B-splines do mesmo grau dentro do mesmo intervalo. Rigby e Stasinopoulos destacam ainda as seguintes características gerais:

- As base spline são definidas por funções locais que têm seu domínio dentro de $2 + D$ nós da faixa x. Por exemplo, para splines cúbicos com $D = 3$ cada função base é definida dentro de 5 nós;

- Os nós não precisam estar na mesma distância, então padrões gerais de nós são possíveis;

- O número de nós determina o tamanho da base B-spline que compõe o função polinomial;

- B-splines são colunas da matriz base B. Esta matriz pode ser usada em uma estrutura de regressão como a matriz de projeto. Os coeficientes de regressão $\widehat{\mathbf{y}} = \mathbf{B}\boldsymbol{\beta}$ produz uma relação não linear flexível entre y e x.

2.3 ANÁLISE DE DIAGNÓSTICO

Para Rigby e Stasinopoulos (2020), [30] , os resíduos simples, de Pearson e de desviance não se comportam muito bem, pois não apresentam distribuição normal para respostas assimétricas. Os autores recomendam para modelos GAMLSS resíduos quantílicos normalizados. O resíduo quantílico é independente da distribuição da variável resposta e apresenta distribuição normal quando o modelo assumido é correto.

Resíduo quantílico foi definido por Dunn e Smyth (1996), [8], como sendo

$$\widehat{r_i} = \Phi^{-1}\left(\widehat{u_i}\right), \tag{2.15}$$

em que $\Phi(.)$ é a função de distribuição acumulada da normal padrão e $\widehat{u_i}$ são resíduos quantílicos, definidos diferentemente para variáveis contínuas e discretas.

Se y é uma observação de uma variável contínua então $u = F(y|\theta)$ e $\widehat{u} = F(y|\widehat{\theta})$ são, respetivamente, resultados do modelo e do ajuste da distribuição acumulada. Dessa forma, pela transformação integral da probabilidade, u tem uma distribuição de uniforme entre 0 e 1.

Se y é uma observação de uma resposta discreta, então $F(y|\theta)$ é uma função escada. Nesse caso, u é definida como um valor aleatório no intervalo $[\widehat{u_1}, \widehat{u_2}] = \left[F(y-1|\widehat{\theta}), F(y|\widehat{\theta})\right]$.

Segundo Stasinopoulos et al (2020), [30], os verdadeiros resíduos seguem uma distribuição normal padrão se o modelo é correto e a distribuição normal tem média 0, variância 1, assimetria 0 e curtose 3.

No R os resíduos quantílicos estimados podem ser obtidos pela função *resid*(). No pacote *gamlss* existem outras funções que usam os resíduos quantílicos, que são: *plot*() e *wp*().

2.3.1 FUNÇÃO PLOT()

A função *plot()* (ou *plot.gamlss()* para ser completo) é usada para avaliação geral dos resíduo. Essa função produz quatro gráficos para verificação dos resíduos quantílicos para um objeto *gamlss* ajustado:

- resíduo versus valores ajustados para o parâmetro μ;
- resíduo versus um índice ou uma covariável especificada;
- estimativa Kernel da densidade dos resíduos;
- QQ-normal plot dos resíduos.

2.3.2 FUNÇÃO WP()

Para Buuren (2007) [31] o worm plot investiga o quão bem o modelo se ajusta aos dados. A função wp() retorna um worm plot único ou múltiplo para objetos gamlss ajustados. O worm plot é um QQ-plot sem tendência. Esta ferramenta de diagnóstico permite checar os resíduos em diferentes regiões de uma ou duas variáveis preditoras.

Os pontos mostram os desvios dos resíduos ordenados de seus valores esperados (aproximados) representados pela curva. Quanto mais próximos dessa linha, mais próxima a distribuição dos resíduos está de uma normal padrão. O intervalo de confiança de 95% dá uma impressão da variação da amostragem e delineia a região onde a minhoca (*worm*) deve ser localizada. A forma da minhoca comunica o tipo de desajuste entre o modelo e os dados. Se os dados forem normais a curva worm-plot deve aparentar uma minhoca achatado, os pontos próximos a curva vermelha e com poucas oscilações. A Figura 2.1 é um exemplo de um worm plot de um regressão, nesse caso todas as observações caem dentro dos limites de confianças e não é detectada uma forma específica nos pontos, no geral o modelo parece se ajustar bem (Buuren (2007) [31]).

Para Rigby e Stasinopoulos (2020), [30] se o modelo estiver correto, espera-se que aproximadamente 95% dos pontos estejam entre as duas curvas elípticas e 5% fora. Para os autores, porcentagem maior de pontos fora as duas curvas elípticas indicam que a distribuição ajustada (ou os termos ajustados) do modelo são inadequados para explicar a variável resposta.

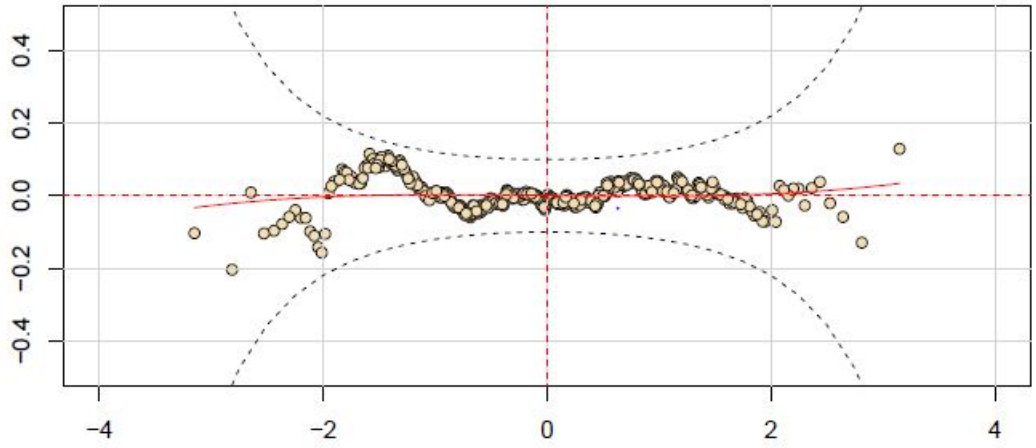

Figura 2.1: Exemplo de worm plot.

Na função do **R**, *wp()* se for especificado uma variável explicativa do modelo (*xvar*) será fornecido *n.iter* gráficos. Neste caso, a variável X é cortada em *n.iter* intervalos não sobrepostos com igual números de observações e os gráficos QQ normais (ou seja, worm) dos resíduos para cada intervalo são plotados.

3 GAMLSS PARA ANÁLISE DE SOBREVIVÊNCIA

Os modelos aditivos generalizados para locação, escala e forma (GAMLSS) são uma poderosa classe de modelos de regressão, capazes de modelar dados que apresentam não-linearidade entre a variável resposta e suas covariáveis. Permitem ainda a estimação dos parâmetros da distribuição de probabilidade da variável resposta, em função das variáveis explicativas. Em análise de sobrevivência a variável aleatória de interesse é o tempo até a ocorrência de um evento de interesse, a partir do acompanhamento do indivíduo (Carvalho et al (2011) [2]). O evento de interesse pode ser a morte, incidência de doença, recaída, recuperação (por exemplo retorno ao trabalho) ou qualquer experiência designada de interesse que pode acontecer com um indivíduo (Kleinbaum et al (2020) [17]).

Na análise de sobrevivência assume-se que T denota a variável aleatória tempo de sobrevivência, isto é, o tempo até a ocorrência do evento de interesse. Essa variável aleatória contínua tem com função densidade de probabilidade $f(t|\boldsymbol{\theta})$, em que $\boldsymbol{\theta}$ é o vetor de parâmetros da distribuição. Nesse contexto, $f(t|\boldsymbol{\theta})$ denota a probabilidade de um indivíduo sobre o evento de interesse em um intervalo instantâneo de tempo, dessa forma para um incremento ε no tempo, resulta na seguinte equação:

$$f(t|\boldsymbol{\theta}) = \lim_{\varepsilon \to 0^+} \frac{P(t \leq T \leq t + \varepsilon)}{\varepsilon}. \tag{3.1}$$

A função de sobrevivência, $S(t|\boldsymbol{\theta})$, é definida como a probabilidade de um indivíduo sobreviver mais do que um determinado tempo t, ou no mínimo um tempo igual a t, portanto, dada por

$$S(t|\boldsymbol{\theta}) = P(Y \geq t) = 1 - F(t|\boldsymbol{\theta}), \tag{3.2}$$

sendo $F(t|\boldsymbol{\theta})$ a função de distribuição acumulada da variável T. Portanto, a função de densidade

de probabilidade da variável T pode ser obtida da seguinte forma

$$f(t|\boldsymbol{\theta}) = -\frac{dS(t|\boldsymbol{\theta})}{dt}. \tag{3.3}$$

A estimação do k-ésimo parâmetro, θ_k, $1 \leq k \leq p$, da distribuição de probabilidade da variável aleatória T pode ser feita pelo Método de Máxima Verossimilhança. Nesse caso, a construção da função de verossimilhança depende do tipo de censura. Segundo Carvalho et al (2011) [2], a censura é à direita quando o evento de interesse ocorre depois do período de observação, enquanto que na censura à esquerda não é conhecido o desfecho, porém sabe-se que ele ocorreu.

Tomando n observações independentes da variável aleatória T identicamente distribuídas com censura à direita. A notação clássica para o i-ésimo indivíduo, $1 \leq i \leq n$, é (T_i, δ_i), em que T_i é o tempo de observação do indivíduo i e δ_i uma variável indicadora de falha. A indicadora δ_i assume o valor 1 se o desfecho ocorreu e 0 se o indivíduo foi censurado. Nessas condições, para o i-ésimo indivíduo a probabilidade de ocorrência do evento no tempo de observação t_i é

$$P_i = \left[f(t_i|\boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i|\boldsymbol{\theta})\right]^{1-\delta_i}. \tag{3.4}$$

Nessas condições, o logaritmo da função de verossimilhança é dado por

$$l(\boldsymbol{\theta}) = \log\left[\prod_{i=1}^{n} P_i\right] = \sum_{i=1}^{n} \delta_i \log\left[f(t_i|\boldsymbol{\theta})\right] + \sum_{i=1}^{n}(1-\delta_i)\log\left[S(t_i|\boldsymbol{\theta})\right], \tag{3.5}$$

e os valores de $\boldsymbol{\theta}$ que maximizam essa função são os parâmetros da função de densidade de probabilidade da variável aleatória T. Dessa forma é possível determinar a Função de Sobrevivência dada pela Equação (3.2).

3.1 MODELOS DE FRAÇÃO DE CURA

Em situações em que a estimação não-paramétricas da Função de Sobrevivência indica a existência de uma assíntota horizontal é possível que haja uma proporção de indivíduos curados. O primeiro trabalho desenvolvido para estimar essa proporção foi desenvolvido por

Berkson e Gagel (1952) [1], que estimou a proporção de curados numa população submetida a um determinado tratamento de câncer. Essa metodologia é frequentemente utilizada quando se deseja analisar indivíduos que são acompanhados por um período longo de tempo e observa-se que uma fração razoável não irá experimentar o evento de interesse (Gonzales (2014) [13]).

Para m causas de risco e q indica uma causa de risco, $1 \leq q \leq m$, então a probabilidade de ocorrência dessa causa é $p_q = P[N = q]$, sendo N a variável aleatória que denota causa de risco. Se não for observado nenhuma causa então denota-se $N = 0$. O tempo até a q-ésima causa ($N = q$) é T_q e o tempo do evento é definido por $T = min\{T_1, T_2, \cdots, T_m\}$. Se $N = 0$ então $T = \infty$.

A função de sobrevivência da variável T é chamada de função de sobrevivência populacional. Para Rodrigues et al (2008) [27] essa função corresponde a probabilidade de nenhum evento de interesse ocorrer até o tempo t, isto é

$$S_{pop}(t) = P(T > t). \tag{3.6}$$

Como $T = min\{T_1, T_2, \cdots, T_m\}$ e $T = \infty$ se $m = 0$, então a Equação (3.6) dado por

$$S_{pop}(t) = P(N = 0) + P(T > t, N \geq 1). \tag{3.7}$$

A função de sobrevivência populacional indica que existe uma proporção da população que não está sujeita a ocorrência do evento de interesse, p_0 (Gonzales (2014) [13]). Dessa forma, a função sobrevivência populacional deve apresentar uma assíntota horizontal em $S_{pop} = p_0$, isto é,

$$\lim_{t \to \infty} S_{pop}(t) = P(N = 0) = p_0. \tag{3.8}$$

Na Equação (3.7), das propriedades da probabilidade condicional, a parcela $P(T > t, N \leq 1)$ pode ser expandida da seguinte forma

$$P(T > t, N \geq 1) = P(T > t | N \geq 1)P(N \geq 1) = \sum_{q=1}^{m} P(T > t | N = q)P(N = q).$$

Considerando que $T = min\{T_1, T_2, \cdots, T_m\}$, pela distribuição do mínimo (James (2015) [16]),

$$P(T > t, N \geq 1) = \sum_{q=1}^{m} \left[\prod_{i=1}^{q} P(T_i > t|N = q) \right] P(N = q) = \sum_{q=1}^{m} \left[\prod_{i=1}^{q} 1 - F(t|N = q) \right] P(N = q).$$

Como as variáveis aleatórias T e N são independentes (Rodrigues et al (2007) [27] e Castro et al (2010) [3]), então

$$\sum_{q=1}^{m} \left[\prod_{i=1}^{q} 1 - F(t|N = q) \right] P(N = q) = \sum_{h=1}^{m} [1 - F(t)]^q \, P(N = q).$$

Como $S(t) = 1 - F(t)$, a Equação (3.7) é reescrita por

$$S_{pop}(t) = P(N = 0) + \sum_{q=1}^{m} [S(t)]^q \, P(N = q). \tag{3.9}$$

Da função de sobrevivência populacional é possível obter a função de densidade de probabilidade populacional. Como na Equação 3.3, é possível obter $f_{pop}(t)$, dada por

$$f_{pop}(t) = -\frac{d}{dt} S_{pop}(t) = \sum_{q=1}^{m} q \, [S(t)]^{q-1} \, f(t) P(N = q). \tag{3.10}$$

3.2 GAMLSS PARA O MODELO WEIBULL DE TAXA DE CURA

No trabalho de Berkson e Gage (1952) [1], foi investigado a taxa de cura de pacientes submetidos a um tratamento de câncer. Nesse caso, em que a variável aleatória N é dicotômica, isto é, $N = 0$ curado e $N = 1$ não curado, em que N é uma variável latente gerada de uma distribuição binomial. A probabilidade do i-ésimo indivíduo ser curado é denotada por $\nu = P(N_i = 0)$, estimado no modelo de fração de cura. Com efeito, a função de sobrevivência populacional (3.9) é reescrita, para o i-ésimo indivíduo temos a seguinte função:

$$S_{pop}(t_i) = \nu + (1 - \nu) S(t_i). \tag{3.11}$$

Por consequência, a função densidade de probabilidade populacional é dada por

$$f_{pop}(t_i) = (1 - \nu) f(t_i), \tag{3.12}$$

na qual a função f é a função densidade de probabilidade da variável T, tal que ν seja a assíntota horizontal da função S_{pop}. Dessa forma f pode ser qualquer distribuição de densidade contínua para $t > 0$.

O modelo Weibull, considera que a densidade de probabilidade da variável T na Equação (3.12) é uma distribuição Weibull. Dessa forma, a função de densidade de probabilidade é dada por

$$f\left(t|(\lambda,\alpha)\right) = \lambda\alpha x^{\alpha-1} \exp\{-\lambda x^{\alpha}\}. \tag{3.13}$$

A média, variância e função de sobrevivência são respectivamente dadas por

$$\mu = \frac{\Gamma(1+\alpha^{-1})}{\lambda^{1/\alpha}} \tag{3.14}$$

$$\sigma^2 = \frac{\Gamma(1+2\alpha^{-1}) - \Gamma^2(1+\alpha^{-1})}{\lambda^{2/\alpha}} \tag{3.15}$$

$$S(t) = \exp\{-\lambda x^{\alpha}\}. \tag{3.16}$$

Reparametrizando a função de distribuição em função de μ e σ, a função de sobrevivência é reescrita por

$$S(t) = \exp\left\{-\left[\frac{t\Gamma(1/\sigma+1)}{\mu}\right]^{\sigma}\right\}. \tag{3.17}$$

Dessa forma, a Equação (3.11) é reescrita da seguinte forma

$$S_{pop}(t|\mu,\sigma,\nu) = \nu + (1-\nu)\exp\left\{-\left[\frac{t\Gamma(1/\sigma+1)}{\mu}\right]^{\sigma}\right\}. \tag{3.18}$$

Ramires et al (2019) [22] apresentou a metodologia de regressão GAMLSS para determinar os parâmetros $\theta = \begin{bmatrix} \mu & \sigma & \nu \end{bmatrix}$ da Equação (3.18). Os autores criaram a função de densidade de probabilidade Weibullcr, modelada pelos parâmetros $\boldsymbol{\theta}$ e fizeram a implementação da mesma no R (https://git.io/vAdIb).

Para implementar o método em questão, através dos modelos de regressão GAMLSS, e

estimar os parâmetros que compõem o vetor $\boldsymbol{\theta}$, tem-se

$$g(\theta_k) = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{h}_{jk}(x_{jk}), \tag{3.19}$$

em que k é o índice dos parâmetros de $\boldsymbol{\theta}$, assim $1 \leq k \leq 3$.

Conforme apresentado do Capítulo 2, sobre modelos de regressão GAMLSS, os parâmetros $\boldsymbol{\beta}$ e os coeficientes $\boldsymbol{\gamma}$ das funções de suavização ($\mathbf{h} = \mathbf{Z}\boldsymbol{\gamma}$) são estimados pela máxima verossimilhança penalizada, assim, temos a seguinte equação

$$l_p = l(\theta) - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \lambda_{jk}\gamma_{jk}^t\mathbf{P}_{jk}\gamma_{jk}, \tag{3.20}$$

sendo $l(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança com censura à direita, dado pela Equação (3.5), e $\mathbf{P}_{jk}$ uma matriz simétrica que pode depender de um vetor de parâmetros de suavização. A estimação dos parâmetros $\theta = \begin{bmatrix} \mu & \sigma & \nu \end{bmatrix}$ depende da função de ligação $g(.)$ conforme apresentado no capítulo anterior. Para o caso do modelo Weibullcr de fração de cura a estimação será definida por

$$\widehat{\boldsymbol{\mu}} = \exp\left[\mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1^{''}} \mathbf{h}_{j1}(\mathbf{x}_{j1})\right], \tag{3.21}$$

$$\widehat{\boldsymbol{\sigma}} = \exp\left[\mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2^{''}} \mathbf{h}_{j2}(\mathbf{x}_{j2})\right], \tag{3.22}$$

$$\widehat{\boldsymbol{\nu}} = \frac{\exp\left[\mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3^{''}} \mathbf{h}_{j3}(\mathbf{x}_{j3})\right]}{1 + \exp\left[\mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{h}_{j3}(\mathbf{x}_{j3})\right]}, \tag{3.23}$$

em que $\mathbf{X}_k$, $1 \leq k \leq 3$ são as matrizes das $J_k^{'}$ covariáveis do correspondente parâmetro k, em que $k = 1$ corresponde à média, $k = 2$ o desvio e $k = 3$ à fração de cura. As matrizes $\mathbf{X}_k$ são de dimensão $n \times J_k^{'}$, assim os coeficientes $\boldsymbol{\beta}_k$, determinados na regressão, são de dimensão $J_k^{'} \times 1$. No modelo, as demais $J_k^{''}$ covariáveis x_{jk}, $1 \leq j \leq J_k^{''}$, são suavizadas por P-splines.

4 ANÁLISE DE SOBREVIVÊNCIA PARA O ESTUDO SOBRE CÂNCER DE MELANOMA

Com o objetivo de avaliar a sobrevivência em pacientes com câncer de melanoma, vamos ajustar o modelo apresentado no Capítulo anterior. Para essa análise, o banco de dados foi obtido no site da Fundação Oncocentro de São Paulo (FOSP)[7], que reúne as informações sobre câncer das instituições de saúde do estado de São Paulo. A FOSP organiza bancos dados sobre o atendimento e tratamento do câncer no referido estado desde o ano 2000. O banco utilizado no presente trabalho reúne os dados a partir de 2014. Foram selecionados os casos de câncer de pele e a descrição das variáveis usadas na análise está na Tabela (4.1).

O estadiamento é uma forma para definir o estágio do câncer a partir de informações recolhidas em diversos testes. O sistema de classificação é o TNM, em que a classificação T fornece detalhes do tumor primário, N detalhes dos linfonodos regionais e M detalha a doença sistêmica (metástases). Para criar a variável estágio seguimos o trabalho de Gomes et al (2017) [12], que está de acordo com a AJCC (American Joint Comission on Cancer), descreve a evolução clinica da doença, mais detalhes estão na Tabelas 6.9 do Apêndice 2. Portanto, cada estágio apresenta a seguinte característica:

- Estágio 0: não foi possível identificar tumor primário;
- Estágio I-II: doença com localização primária;
- Estágio III: doença com localização na área loco regional;
- Estágio IV: doença generalizada.

O câncer de pele apresenta alta taxa de cura, segundo Silversmit et al (2016) [29] a proporção de cura é de 81%, para mulheres a fração de cura é de 84% e homens 75%. Por essa informação, torna-se necessário a verificação da existência da fração de cura no dados coletados. O desfecho de interesse nesta análise é a morte pelo câncer e os demais desfecho são

Tabela 4.1: Descrição das variáveis utilizadas na análise.

Variável		Amostra Total
Total da amostra (n)		14324
Sexo		
	Masculino	7568 (52,83%)
	Feminino	6756 (47,16%)
Idade		$\overline{X}$ = 68,77 anos (S = 13,86 anos)
Tratamento		
	Cirurgia	12864 (89,8%)
	Outras combinações de Tratamento	573 (4%)
	Quimioterapia e/ou Radioterapia	462 (3,22%)
	Nenhum	425 (2,96%)
tempo		
	Data Ultima Info - Data Diagnostico	$\overline{X}$ = 2,65 anos (S = 2,34 anos)
Última Informação		
	Vivo curado	10721 (74,84%)
	Óbito por outras causas	2671 (18,64%)
	Vivo com Câncer	578 (4,03%)
	Óbito por Câncer	354 (2,47%)
Categoria de Atendimento		
	SUS	11219(78,32%)
	Não SUS	3105(21,67%)
Clinica Médica		
	Dermatologia	5585(38,99%)
	Oncologia clínica	3175(22,16%)
	Outras	2657(18,55%)
	Oncologia Cutânea	1746(12,19%)
	Cirurgia Oncológica	1139(8,19%)
Estágio do Câncer		
	0	1367 (9,54%)
	I	10914 (76,19%)
	II	1748 (12,20%)
	III	168 (1,17%)
	IV	127 (0,88%)

definidos como censura. Com essa finalidade, foi feita a estimação da função de sobrevivência pelo método de Kaplan-Meier. O resultado é apresentado na Figura (4.1) e indica a existência de uma assíntota horizontal, indicando a existência de uma fração de cura. O Teste de Maller e Zhou (1992) [19] indica que a assíntota horizontal observada é estatisticamente significativa com $p.valor < 0{,}001$.

Calculando a função de sobrevivência pelo método de Kaplan-Meier para as variáveis Sexo, Categoria de Atendimento, Estágio e Tratamento também é verificado a existência de assíntotas horizontais, como indica a Figura(4.2). Nota-se que há diferença entre as curvas de sobrevivência de homens e mulheres e que quanto maior o número do estágio menor a perspectiva de cura. Quanto à categoria de atendimento, pacientes atendidos por convênio ou particular (Não SUS) apresentam melhor perspectiva de cura.

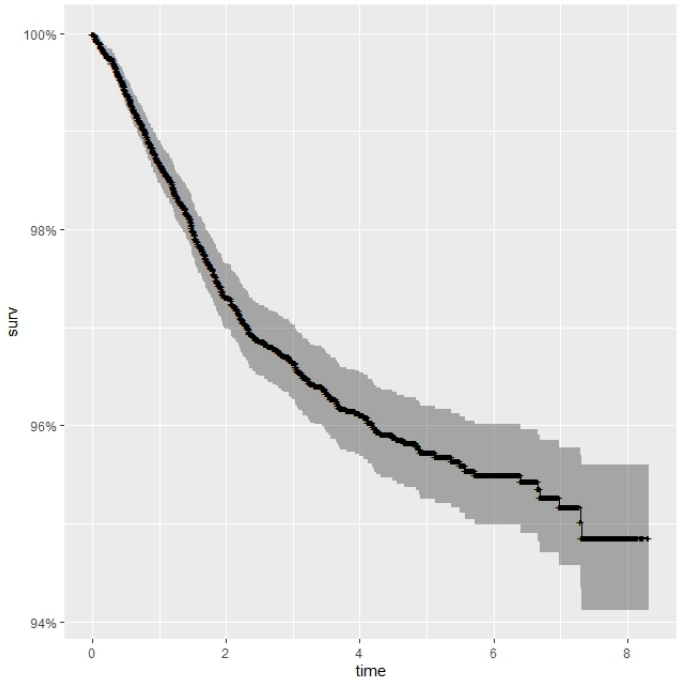

Figura 4.1: Evolução da função de sobrevivência ao longo do tempo (em anos) obtido pelo método de Kaplan-Meier.

Em todos os casos verificados pelo método de Kaplan-Meier, a função de sobrevivência não vai para zero e sim para um platô, além disso a proporcionalidade dos riscos não é mantida, como verificado no cruzamento das curvas de sobrevivência quanto ao Estágio da doença na Figura (4.2D). Por esses motivos o modelo de Cox não é recomendável para o problema em questão.

Com a finalidade de modelar a fração de cura do câncer de pele foi considerado o modelo de regressão Weibull proposto por Ramires et al (2019) [22], denominado por regressão Weibullcr. A seleção das variáveis que entraram no modelo de regressão foi feita pelo critério de Akaike (AIC). Primeiro foi ajustado a média, deixando os demais parâmetros sem covariáveis até a obter o melhor AIC. O processo foi repetido para os demais parâmetros e a relação de variáveis para modelagem de cada parâmetro está descrita na Tabela (4.2). No modelo, para ajustar os parâmetros μ foi utilizado a função de suavização P-Spline na variável Idade.

A Tabela (4.3) apresenta os valores estimados para os parâmetros $\boldsymbol{\beta}$, o Erro Padrão (EP) e o p-valor do ajuste ao modelo de regressão Weibullcr para os parâmetros de Localização (μ), Dispersão (σ) e Fração de Cura (ν), respectivamente. Para os termos de suavização presente

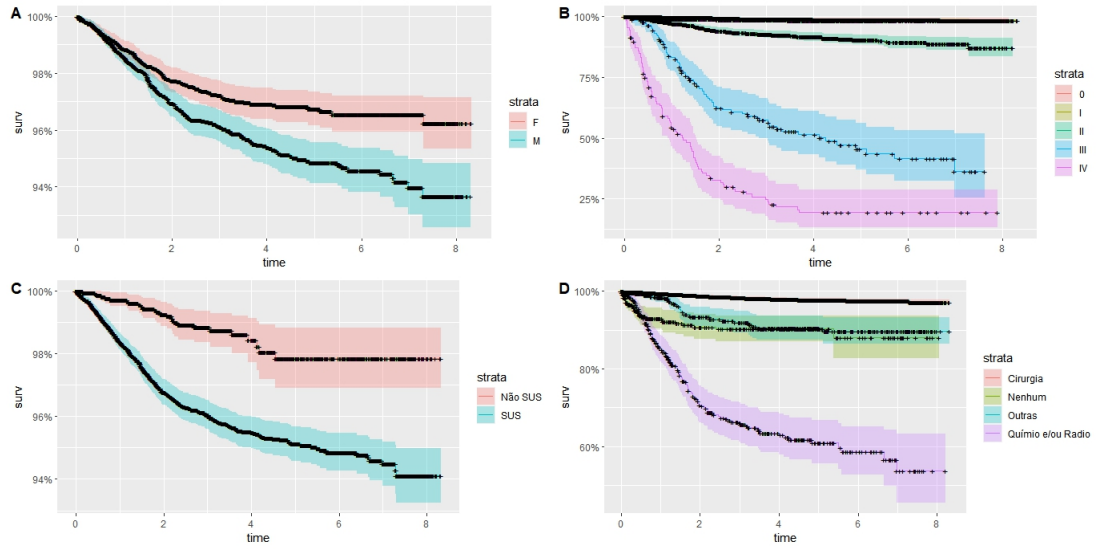

Figura 4.2: Evolução da função de sobrevivência ao longo do tempo (em anos) obtido pelo método de Kaplan-Meier para as variáveis: Sexo(A), Estágio(B), Categoria de Atendimento (C) e Tratamento (D)

.

Tabela 4.2: Variáveis utilizadas em cada parâmetro.

Parâmetro	Ligação	Variáveis
μ	logaritmo	ps(Idade) + Tratamento + Tipo de Atendimento + Estágio
σ	logaritmo	Tratamento + Estágio
ν	logística	Tratamento + Estágio + Tipo de Atendimento + Sexo

no ajuste do parâmetro (μ), é apresentado na Tabela (4.3) o grau de liberdade (df) e o valor do parâmetro de suavização λ, presente na Equação (3.20). O modelo apresentou um $AIC = 2527$ e não foram significativos os coeficientes das seguintes variáveis:

- Tratamento para a categoria Outros Tratamentos no ajuste do parâmetro σ;
- Estágio para a categoria I no ajuste do parâmetro ν.

Os efeitos parciais das variáveis explicativas Idade modelados usando splines penalizadas, na regressão do parâmetro μ é exibido na Figura (4.3). Por esse resultado, verifica-se para o parâmetro de locação, μ, que a média de sobrevivência é máxima em torno de 50 anos. Algumas conclusões podem ser retiradas para cada um dos parâmetros estimados na Tabela (4.3).

μ: O parâmetro de localização μ indica que o tipo de tratamento, o tipo de atendimento e o estágio diminuem a média do logaritmo do tempo de vida em pacientes suscetíveis $E(log(T)|N = 1) = \log(\mu)$. Dessa forma, o paciente que não está submetido a nenhum

Tabela 4.3: Estimação dos coeficientes de ajuste para os parâmetros μ, σ e ν.

μ	β	EP	p-valor
Intercepto	6,8646	0,05049	$< 2 \times 10^{-16}$
ps(Idade)	$df = 12$	$\lambda = 9750{,}72$	–
Trat = Nenhum	−1,5928	0,06633	$< 2 \times 10^{-16}$
Trat = Outro	−0,3531	0,04144	$< 2 \times 10^{-16}$
Trat = Quimio e/ou Radio	−0,4767	0,05973	$1{,}56 \times 10^{-15}$
Atendimento = SUS	−0,3684	0,04965	$< 2 \times 10^{-16}$
Estágio = I	−0,3789	0,03175	$< 2 \times 10^{-16}$
Estágio = II	−1,6343	0,03838	$< 2 \times 10^{-16}$
Estágio = III	−4,5574	0,06095	$< 2 \times 10^{-16}$
Estágio = IV	−4,1914	0,0796	$< 2 \times 10^{-16}$
σ			
Intercepto	−0,09317	0,02127	$1{,}20 \times 10^{-5}$
Trat = Nenhum	−0,38146	0,03794	$< 2 \times 10^{-16}$
Trat = Outro	−0,38146	0,03336	0,138
Trat = Quimio e/ou Radio	−0,39928	0,03981	$< 2 \times 10^{-16}$
Estagio = I	0,16537	0,02235	$1{,}45 \times 10^{-13}$
Estagio = II	0,16119	0,02822	$1{,}14 \times 10^{-8}$
Estagio = III	0,66924	0,06457	$< 2 \times 10^{-16}$
Estagio = IV	0,58936	0,07519	$4{,}87 \times 10^{-15}$
ν			
Intercepto	1,34319	0,05337	$< 2 \times 10^{-16}$
Trat = Nenhum	1,11268	0,07474	$< 2 \times 10^{-16}$
Trat = Outro	−0,50621	0,11994	$2{,}45 \times 10^{-5}$
Trat = Quimio e/ou Radio	−0,41227	0,15693	0,00862
Estagio = I	−0,08759	0,05106	0,08629
Estagio = II	−3,21623	0,47379	$1{,}18 \times 10^{-11}$
Estagio = III	−0,41891	0,20532	0,04134
Estagio = IV	−0,12165	0,41561	0,0026
Sexo = Masculino	−0,60065	0,03474	$< 2 \times 10^{-16}$
Atendimento = SUS	−1,17031	0,03527	$< 2 \times 10^{-16}$

tratamento diminui a média do logaritmo do tempo de vida em −1,5928 por unidade em relação à cirurgia. Paciente atendido pelo SUS diminui −0,3684 por unidades quando comparamos em relação ao Não-SUS. Quanto ao estágio, verifica-se que quanto mais avançado for o estágio da doença, maior será a diminuição do logaritmo do tempo de vida dos paciente.

σ: O parâmetros de dispersão σ indica que o tipo de tratamento diminui a variabilidade dos tempos de falha em escala logarítmica ($\log(\sigma)$), de forma que pacientes que recebem

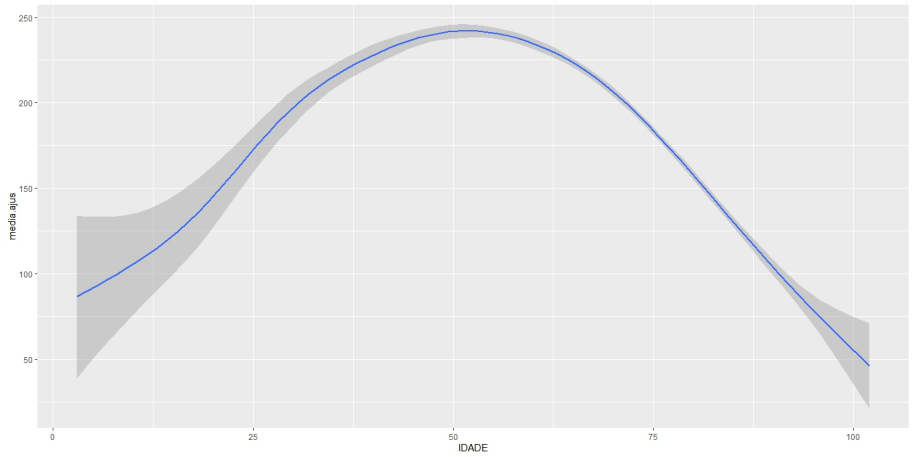

Figura 4.3: Ajuste não linear para o parâmetro de localização μ em relação a Idade.

Quimioterapia e/ou Radioterapia apresentam a maior redução da variabilidade, −0,3993 por unidade em relação à cirurgia. Quanto ao estágio da doença, quanto mais avançado o câncer, maior o crescimento de variabilidade no tempo de falha. Porém, pacientes no estágio III apresentam maior aumento na variabilidade do tempo de falha, 0,6692 por unidade em relação ao estágio 0.

ν: O parâmetro de fração de cura ν indica que pacientes que não receberam nenhum tratamento tem chance de cura 3,03 ($e^{1,11268}$) vezes maior do que pacientes que fizeram cirurgia. Já pacientes que receberam Outros tratamentos ou fizeram Quimioterapia e/ou Radioterapia apresentam chance menor de cura em relação ao pacientes que foram submetidos à cirurgia (0,603 e 0,662, respectivamente, vezes a chance de quem fez cirurgia). Pacientes que estão no Estágio II da doença apresentam menor chance de relação ao pacientes que estão no Estágio 0 (0,04 vezes a chance de cura no Estágio 0). Os demais Estágio também apresentam menor chance de cura em relação ao Estágio 0. Pacientes homens e pacientes atendidos pelo SUS também apresentam menor chance cura em relação à mulheres e pacientes não SUS, respectivamente.

A estimativa da fração de cura para os diferentes estratos das covariáveis categóricas é apresentada na Figura (4.4). Por essa estimação, verifica-se que homens apresentam menor fração de cura em relação as mulheres; o Estágio II da doença apresentou a menor fração de cura comparado com os demais estágios, enquanto que o Estágio 0 apresentou maior fração de cura; os

pacientes que receberam nenhum tratamento apresentaram maior fração de cura em relação aos que receberam.

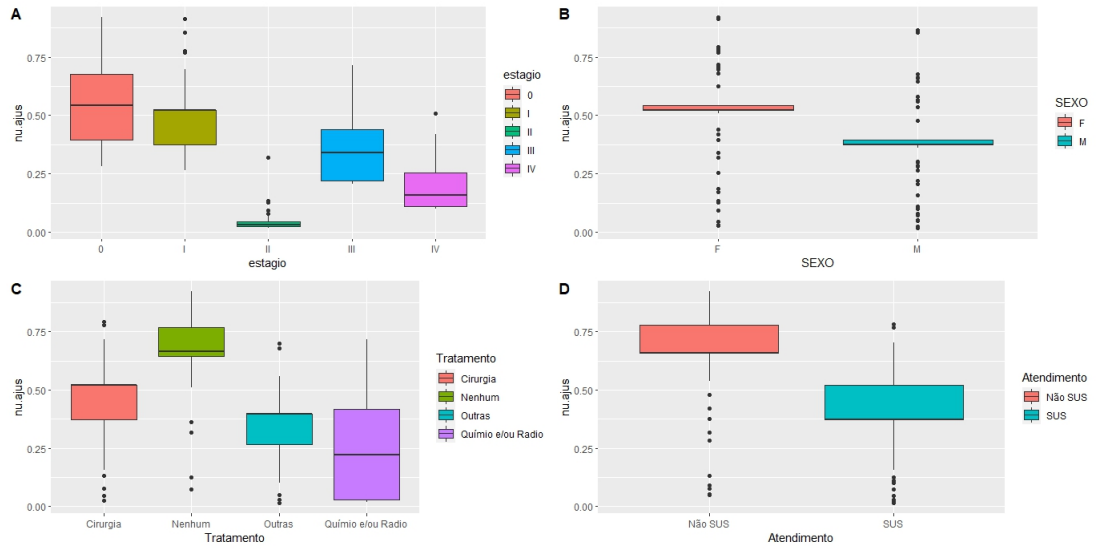

Figura 4.4: Estimativa das Fração de Cura (ν) pelo regressão GAMLSS em cada categoria.

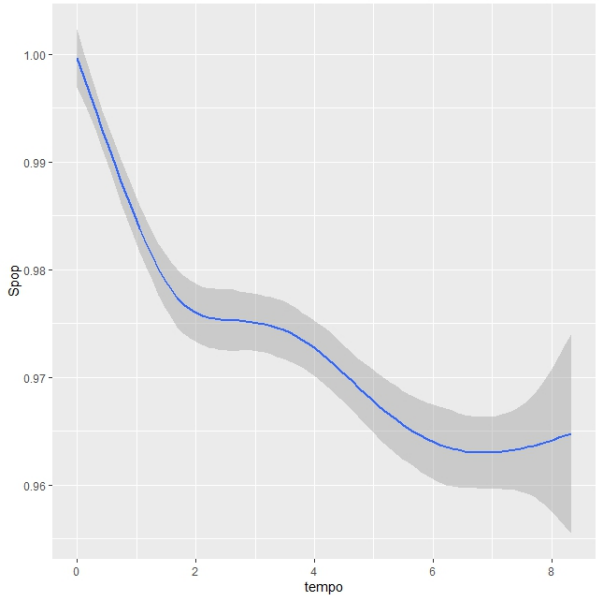

Figura 4.5: Função de Sobrevivência da População ajustada.

A função de sobrevivência da população $S_{pop}(.)$ definida na Equação (3.18) foi estimada a partir dos valores ajustados e o gráfico é apresentado na Figura (4.5). Por esse resultado verifica-se a possível existência de uma assíntota horizontal maior de 0,96 e menor que 0,97. A Figura (4.6) mostra a relação entre a função de sobrevivência da população com cada uma

das variáveis e categorias. Verifica-se uma probabilidade maior de sobrevivência para pacientes que não foram atendidos pelo SUS. Mulheres apresentam maior probabilidade de sobrevivência que homens em quase todo período. Quanto ao estágio da doença, pacientes nos Estágios 0 e I apresentam praticamente a mesma probabilidade de cura, superior à 0,9. A Figura (4.6) ainda indica que o avanço dos estágios da doença implica na diminuição da probabilidade de cura. Pacientes no Estágio IV apresentam probabilidade de cura em torno de 0,25. No que diz respeito ao Tratamento, pela Figura (4.6), é possível verificar que a cirurgia apresenta melhor fração de cura, superior a 0,95, seguido de Outros Tratamentos. Pacientes que não recebem tratamento tendem a ter sua expectativa de cura diminuída após 5 anos do diagnóstico. Pacientes submetidos à tratamentos com Quimioterapia ou Radioterapia apresentam taxa de cura em torno de 0,7.

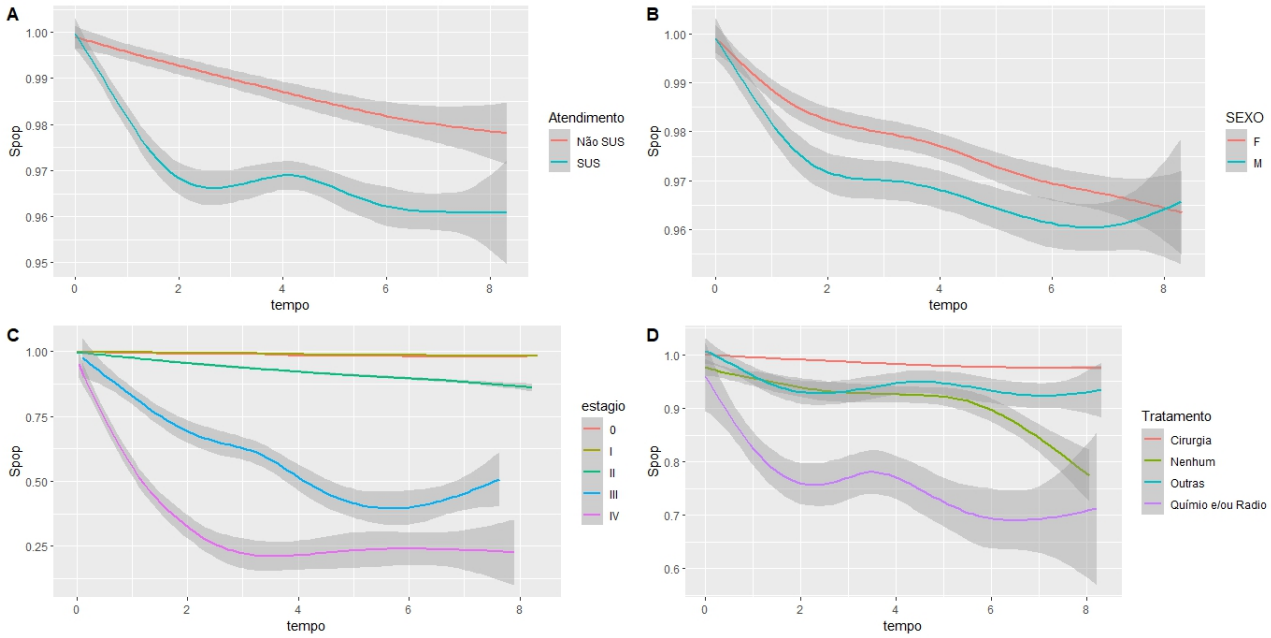

Figura 4.6: Função de Sobrevivência da População ajustada por variável e categoria.

A Figura (4.7) exibe gráficos dos resíduos com a finalidade de verificar a adequação e suposição do modelo proposto. No painel (a) é apresentado o worn plot do ajuste, que indica que não há evidencias de inadequação no modelo, uma vez que a grande maioria dos resíduos caem na região de aceitação dentro das curvas elípticas. Os painéis (b) e (c) apresentam a distribuição e o qqplot dos resíduos quantílicos. Os gráficos indicam que há evidência de que os resíduos quantílicos são normalmente distribuídos com média 1 e variância 0. A normalidade dos resí-

duos também é confirmada pelo teste de correlação de Filleben, que não rejeita a hipótese nula de que os resíduos são normalmente distribuídos, com $p - valor = 0{,}99$.

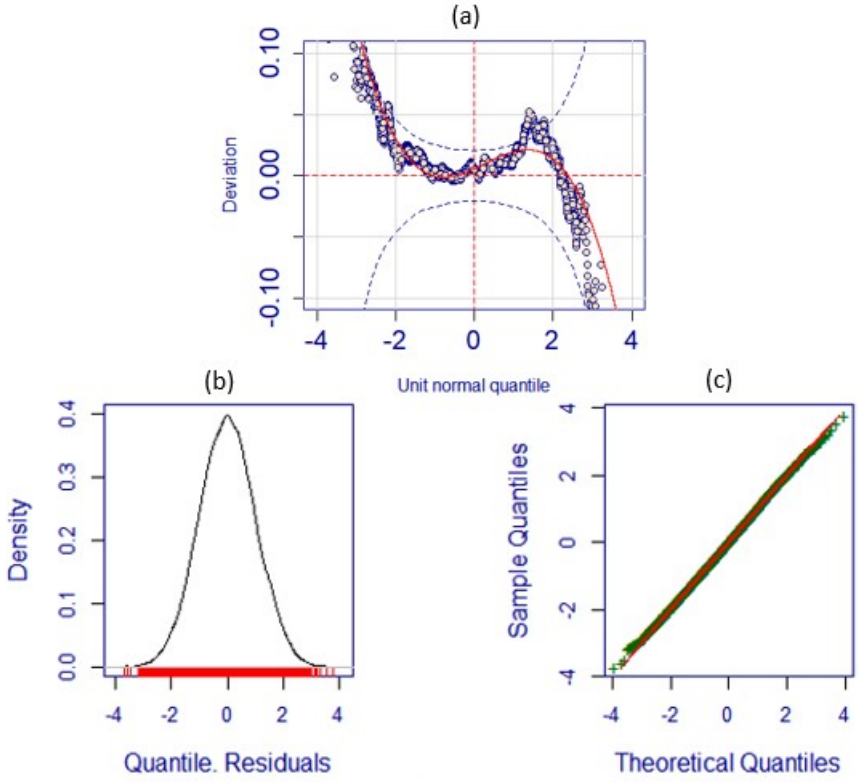

Figura 4.7: Analise de Resíduos: (a) Worm Plot (b) Distribuição dos Resíduos Quantílicos (c) QQPlot dos resíduos quantílicos

5 CONCLUSÕES E PERSPETIVAS

O presente trabalho apresentou uma singela explanação sobre a construção dos modelos de regressão GAMLSS e alguns exemplos, que são apresentados no Apêndice. O GAMLSS surge da necessidade de obter modelos de regressão para dados que seguem distribuições de probabilidade além da Família Exponencial. Tal metodologia ainda permite ajustar os parâmetros de Locação (μ), Escala (σ) e Forma (ν e τ) em função de variáveis independentes do problema em análise (covariáveis). O GAMLSS possibilita tratar dados com comportamento não-linear, permite a implementação de funções de suavização nas covariáveis para melhorar o desempenho do ajuste.

O trabalho retomou alguns conceitos de Análise de Sobrevivência, com foco na aplicação do modelo de regressão GAMLSS para Análise de Sobrevivência proposto por Ramires (2019) [22]. Esse método permite modelar a fração de cura, isto é, a probabilidade de ser curado. Neste método foi utilizado o modelo de mistura, que permitiu agregar a fração de cura como um parâmetro da distribuição.

Para mostrar a aplicabilidade da metodologia GAMLSS para Análise de Sobrevivência, foi utilizado o modelo proposto por Ramires et al (2019) [22] para estimar a fração de cura de pacientes com câncer de melanoma do estado de São Paulo. Foi considerado o modelo de mistura em que o tempo tem distribuição Weibull.

De acordo com a estimativa da fração de cura apresentado na Figura (4.4), pacientes atendidos pelo SUS tem menor fração de cura e portanto menor chance de cura. Quando avaliado por sexo, homens apresentam menor chance de cura que mulheres. Quanto aos diferentes tipos de Tratamentos, pacientes que foram submetidos à Radioterapia e/ou Quimioterapia apresentaram menor fração de cura, que os demais pacientes submetidos à cirurgia. Pacientes que não receberam nenhum tratamento apresentaram maior fração de cura do que pacientes submetidos à cirurgia. Tal fato pode ser reflexo do tempo de acompanhamento do paciente, pois pacientes

que não receberam nenhum tratamento apresentaram menor tempo de acompanhamento (2,35 anos em média).

Conforme o mesmo resultado, verificou-se que o Estágio II apresenta menor fração de cura em relação ao Estágio III. Essa diferença é acentuada nos últimos anos de acompanhamento, quando se verifica maior fração de cura para o Estágio III em relação ao Estagio II. A diferença de fração de cura nos Estágios pode ser explicada pelo Tratamento, pois pacientes que receberam Quimioterapia ou Radioterapia apresentam menor probabilidade de cura. Conforme o banco, o número de pacientes no Estágio II (140 pacientes) que faz Quimioterapia ou Radioterapia é praticamente o dobro dos número de pacientes no Estágio III (71 pacientes).

A regressão GAMLSS se mostrou satisfatória, dada a flexibilidade de modelagem que possibilitou ajustar o modelo de sobrevivência Weibull com fração de cura. Essa metodologia pressupõem uma distribuição que não pertence à Família Exponencial. Para trabalhos futuros, com a finalidade de expandir as possibilidades de modelagem, a presente investigação sugere:

- desenvolver e implementar modelos de regressão GAMLSS com outras distribuições de probabilidade para modelos de mistura;
- implementar a metologia desenvolvida para determinar a fração de cura em outros casos possíveis;
- incluir outras variáveis ao modelo da aplicação, com auxílio de pesquisadores da área, para refinar os resultados.

6 APÊNDICE

6.1 APÊNDICE 1: IMPLEMENTAÇÃO NO R

Na implementação do GAMLSS em R, a função *gamlss()* do pacote gamlss permite modelar todos os parâmetros da distribuição de forma linear e/ou não linear e/ou 'não paramétrico' com funções de suavização das variáveis explicativas.

A função *gamalss()* é semelhante às funções *gam()* no **R** e pode modelar todos os parâmetros da distribuição como função das variáveis explicativas. Pode ser usada para ajustar modelos que podem ser ajustados pelo função *glm()*.

Na sequência são apresentados alguns exemplos da implementação do modelo GAMLSS usando a função *gamlss()* no R.

6.1.1 EXEMPLO 1: DADOS DE INTERNAÇÃO

Exemplo proposto por Rigby e Stasinopoulos (2020), [30] cujos dados estão disponíveis no pacote *gamlss* pelo comando **data(aep)**. O banco de dados do exemplo é composto de 1383 observações de um estudo no Hospital del Mar, Barcelona, durante os anos de 1988 e 1990, apresentado por Gande et al (1996) [11]. A variável de resposta é o número de dias inapropriados (noinap) do total de dias (los) que os pacientes passaram no hospital. Cada paciente foi avaliado quanto à permanência inadequada em cada dia por dois médicos que usaram o protocolo de avaliação de adequação (AEP) [11]. A pesquisa tem por objetivo demonstrar o ajuste de uma distribuição beta binomial aos dados. As variáveis da pequisa são:

los : número total de dias;

loglos : o logaritmo decimal de los;

noinap : número de dias inapropriados de permanência do paciente no hospital;

sexo : sexo do paciente;

enf : tipo de enfermaria do hospital (médico, cirúrgico ou outro);

ano : 1988 ou 1990;

idade : idade do paciente subtraída de 55

y : noinap

Gande et al (1996) [11] comparou os modelos de regressão logística com o modelo de regressão beta binomiais. Os autores verificaram que a regressão beta binomial é um poderosa ferramenta em pesquisa em saúde. Para Gande O uso de modelos beta binomiais não só permite avaliar diferentes probabilidades de acordo com as covariáveis, mas também permite estimar o grau de agrupamento. Nesse trabalho foi modelado a média e a dispersão da distribuição beta binomial em função das variáveis explicativas usando o método epidemiológico EGRET, o que permitiu obter um modelo usando um link logit para a média e um link identidade para a dispersão.

O objetivo do exemplo é melhorar o modelo usando o GAMLSS. Os modelos propostos por Rigby e Stasinopoulos (2020), [30] são:

Modelo I

$$\begin{cases} logit(\mu) = 1 + enf + loglos + ano \\ log(\sigma) = 1 + ano \end{cases}$$

Modelo II

$$\begin{cases} logit(\mu) = 1 + enf + loglos + ano \\ log(\sigma) = 1 + ano + enf \end{cases}$$

Modelo III

$$\begin{cases} logit(\mu) = 1 + enf + cs(loglos,2) + ano \\ log(\sigma) = 1 + ano + enf \end{cases}$$

Modelo IV

$$\begin{cases} logit(\mu) = 1 + enf + cs(loglos,2) + ano + cs(idade,2) \\ log(\sigma) = 1 + ano + enf \end{cases}$$

Nota-se que os Modelos III e IV apresentam suavizações para algumas variáveis. O comando **cs()** indica um spline cúbico. No caso em questão são 2 graus de liberdade. Para as simulações foi considerado que a variável resposta segue a distribuição Beta Binomial, BB(). Para Rigby e Stasinopoulos (2020), [30] a escolha do modelo se dá pelos AIC e BIC.

Tabela 6.1: Valores de AIC e BIC calculados dos Modelos propospos por Rigby e Stasinopoulos (2020), [30] para os dados do problema apresentado por Gande et al (1996) [11]
.

Modelo	AIC	BIC
I	4533,441	4570,065
II	4501,020	4548,108
III	4479,427	4541,147
IV	4454,362	4531,75

As funções ajustadas para todos os termos da média do Modelo IV são mostradas na Figura (6.1) e para os termos do desvio na Figura (6.2).

Para finalizar a Figura(6.3) apresenta o worm plot. Por essa análise verifica-se que a maioria dos resíduos estão no intervalo limitado pelas curvas elípticas além disso, dada a oscilação dos pontos não é possível que os resíduos sejam normalmente padronizados.

6.1.2 EXEMPLO 2: CÉLULAS CD4

Exemplo proposto por Rigby e Stasinopoulos (2020), [30] cujos dados estão disponíveis no pacote *mass* pelo comando **data(cd4)**. O banco de dados do exemplo é composto de 609 observações de contagem de células CD4 em crianças não infectadas com HIV de mães soropositivo feitas por Wade e Ades (1994) [32]. O banco é composto apenas por duas variáveis:

cd4 : contagem de células cd4 em crianças não infectadas com HIV de mães soropositivo;

idade : idade da criança em anos.

Tradicionalmente, problemas desse tipo eram tratados por uma transformação na variável resposta ou uma transformação tanto na resposta quanto na(s) variável(is) explicativa(s). Para simular a não linearidade Rigby e Stasinopoulos (2020), [30] propuseram ajustar polinômios ortogonais de diferentes ordens aos dados e escolher o melhor usando um critério AIC e BIC.

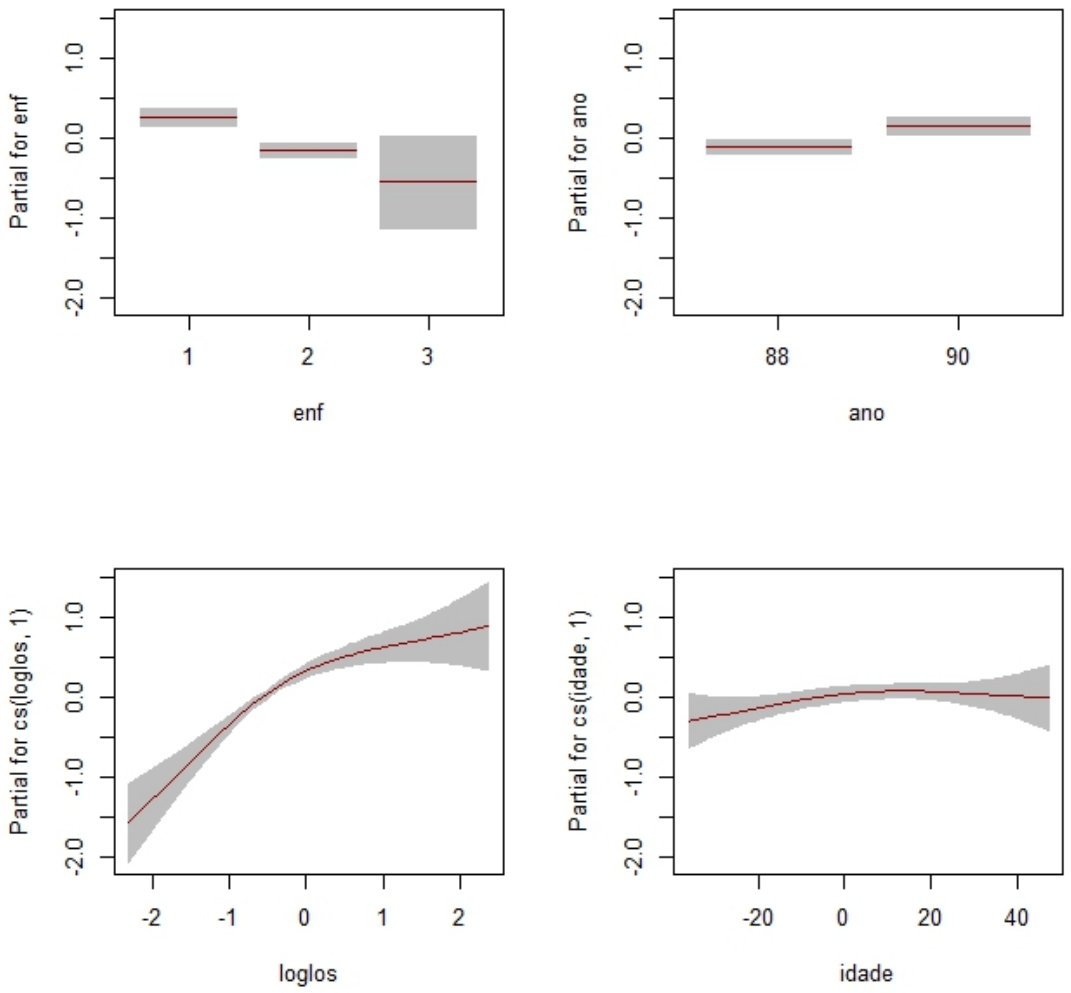

Figura 6.1: Os termos ajustados para média no Modelo 4

Rigby e Stasinopoulos (2020), [30] consideraram ainda que $cd4 \sim N(\mu,\sigma)$ sendo σ constante. A Tabela (6.2) apresenta os valores de AIC e BIC para os modelos ajustados. O número do modelo corresponde ao grau do polinômio. Conforme os valores observados, conclui-se que o modelo 7, isto é o modelo ajustado por polinômio ortogonal de 7° grau, apresenta melhores indicadores. Na Figura (6.4) é apresentada a comparação entre a curva ajusta pelo modelo com polinômio ortogonal de 7° grau.

Como o exemplo apresenta apenas uma variável explicativa então é possível segmentar o intervalo em *n.inter* intervalos com o mesmo número de observações e avaliar a aderência do

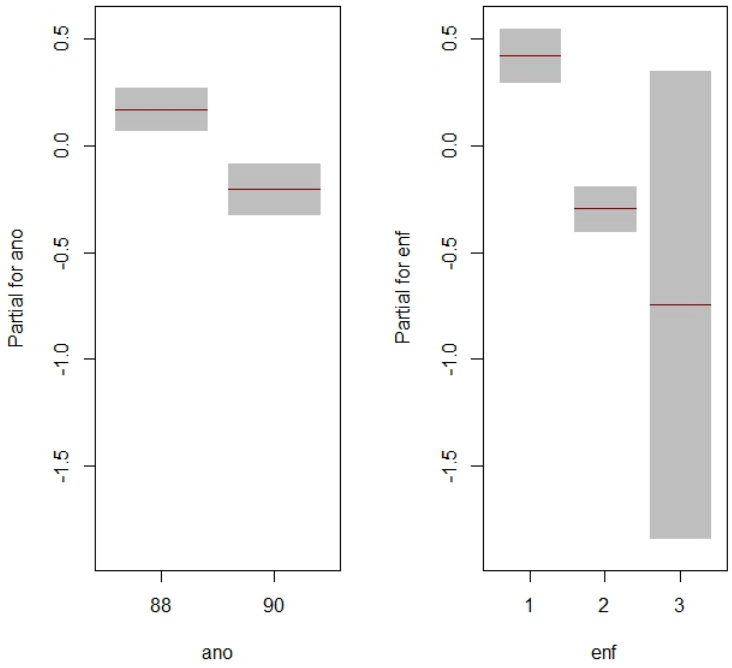

Figura 6.2: Os termos ajustados para o desvio no Modelo 4

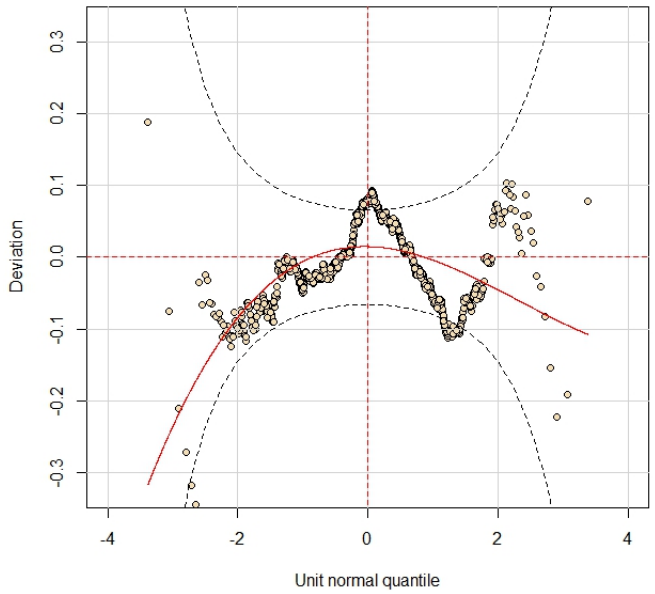

Figura 6.3: O worm plot do modelo 4

ajuste em cada intervalo. Por exemplo o comando

wp(m7,xvar=CD4$ age,n.inter=9)

Tabela 6.2: Valores de AIC e BIC calculados dos Modelos propostos por Rigby e Stasinopoulos (2020), [30] para os dados do problema apresentado por Wade e Ades (1994) [32]. A coluna Grau corresponde ao grau do polinômio ortogonal do ajuste.

Modelo	Grau	AIC	BIC
VII	7	8963,263	9002,969
VIII	8	8963,874	9003,932
VI	6	8968,637	9007,992
V	5	8977,383	9008,266
IV	4	8988,105	9013,284
III	3	8993,351	9014,576
II	2	8995,636	9015,410
I	1	9044,145	9054,380

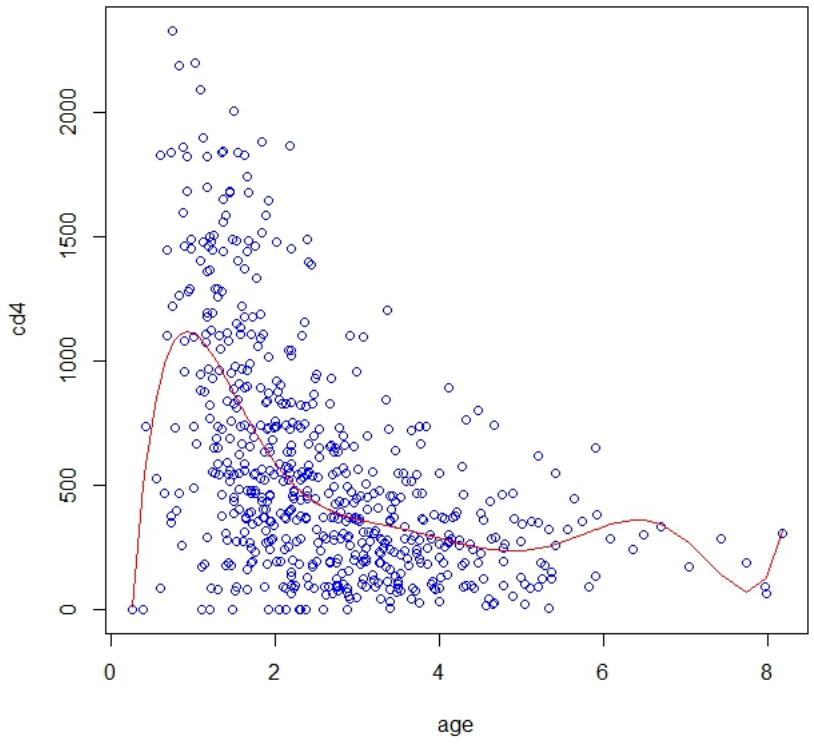

Figura 6.4: Comparação entre a curva ajustada e os dados do Exemplo 2

divide a variável explicativa em 9 subintervalos com aproximadamente o mesmo número de observações e gera os worms plot de cada intervalo. Os gráficos para cada um dos intervalos está na Figura (6.5). A leitura é feita da esquerda para a direita, de baixo para cima. Assim o gráfico de baixo da coluna da esquerda corresponde ao primeiro subintervalo. Nota-se que a curva ajustada melhor adere aos dados nos intervalo no 4° e 5° subintervalo. Na Figura 6.5 nota-

se que a variância pode ser dependente da variável explicativa, pois há maior variabilidade entre 0 e 4. Rigby e Stasinopoulos (2020), [30] propuseram ainda ajuste com polinômio de expoente fracionário usando o comando *gamlss*, supondo dispersão constante. Porém os resultados são semelhantes aos já observados no ajuste por polinômios ortogonais. A presente análise propõe

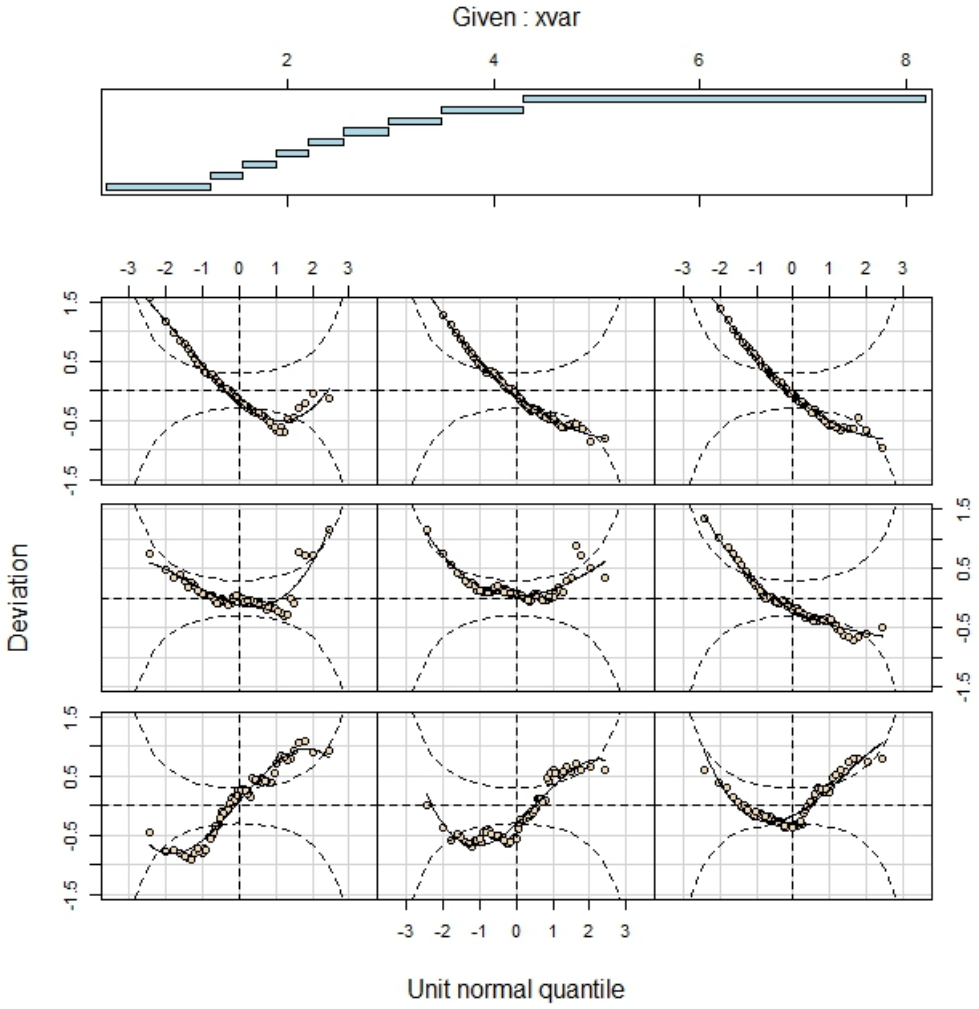

Figura 6.5: Comparação entre a curva ajustada e os dados do Exemplo 2.

o ajuste com expoente fracionário para a média e um ajuste linear para a dispersão. O ajuste por polinômio fracionário é implementado na variável explicativa de interesse x pelo comando *fp(x, n)*, sendo n é o número de potências do ajuste assim,

$$\mu = \beta_0 + \sum_{i=1}^{n} \beta_i x^{p_i}$$

Para o exemplo foram testados três modelos que estão descritos na Tabela (6.3), que indica que

Tabela 6.3: Valores de AIC e BIC calculados dos Modelos com expoente fracionário. A coluna *n* corresponde ao número de potência do ajuste.

Modelo	n	AIC	BIC	Desviance
III	3	8807,174	8846,881	8798,174
II	2	8814,606	8845,489	8807,606
I	1	8836,894	8858,954	8831,894

o Modelo III, isto é com $n = 3$ potências no ajuste, apresenta melhores indicadores. Assim

$$\mu = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2} + \beta_3 x^{p_3}.$$

Os valores de β são obtidos pelo comando *modelo$ mu.coefSmo* e as potências pelo comando *getSmo(modelo)$ power*. No ajuste proposto para duas potências são iguais a −0,5, segundo Rigby e Stasinopoulos (2020), [30], nesse caso o ajuste proposto para a média

$$\widehat{\mu} = 3137{,}08 - 7013x^{-1} + 4919x^{-1/2} - 5179x^{-1/2}log(x)$$

Para o parâmetro de dispersão σ para a distribuição normal tem-se que $\sigma = e^{\eta}$ portanto a equação do ajuste para a dispersão é

$$\sigma = e^{6{,}56-0{,}291x}$$

O representação gráfico do ajuste dos parâmetros e suas respectivas bandas de confiança são apresentados na Figura (6.6). O worm plot para a variável explicativa segmentada em 9 partes é apresentado na Figura (6.6). Verifica-se que a adequação dos resíduos em todos os intervalos, indicando que o modelo é melhor ajustado se comparado com o ajuste por polinômio de 7° grau.

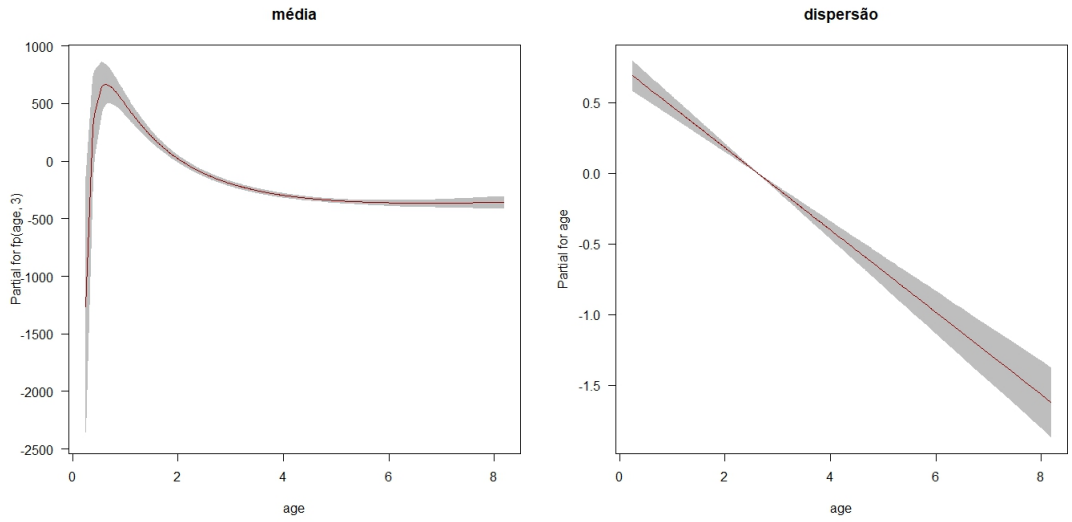

Figura 6.6: Worm plot por segmento do ajuste pelo polinômio fracionário para a média e linear para a dispersão.

6.2 APÊNDICE 2: IMPLEMENTAÇÃO DO GAMLSS PARA SOBREVIVÊNCIA

6.2.1 EXEMPLO 4

Para verificar a eficácia do pacote **gamlss.cens** e comparar com os procedimentos já consolidados, foi replicado o exemplo apresentado por Carvalho et al (2011) [2]. O exemplo proposto tem por objetivo investigar o efeito do tratamento, controlado do HIV por idade e sexo. O banco de dados é proveniente de coorte constituída de pacientes portadores de HIV atendidos entre 1986 e 2000 no Instituto de Pesquisas Clínica Evandro Chagas (Ipec/Fiocruz). Dessa coorte, obteve-se uma amostra de 193 indivíduos que foram diagnosticados com AIDS (critério CDC 1993) durante o período de acompanhamento.

Variáveis:

- tempo: dias de sobrevivência do diagnóstico até o óbito;
- sexo: F (Feminino) e M (Masculino);
- idade: idade na data do diagnóstico;
- tratam: terapia antirretroviral: 0 = nenhuma, 1 = mono, 2 = combinada, 3 = potente.

A relação das variáveis sexo, idade e tratam com o tempo é apresentada na Figura (6.8).

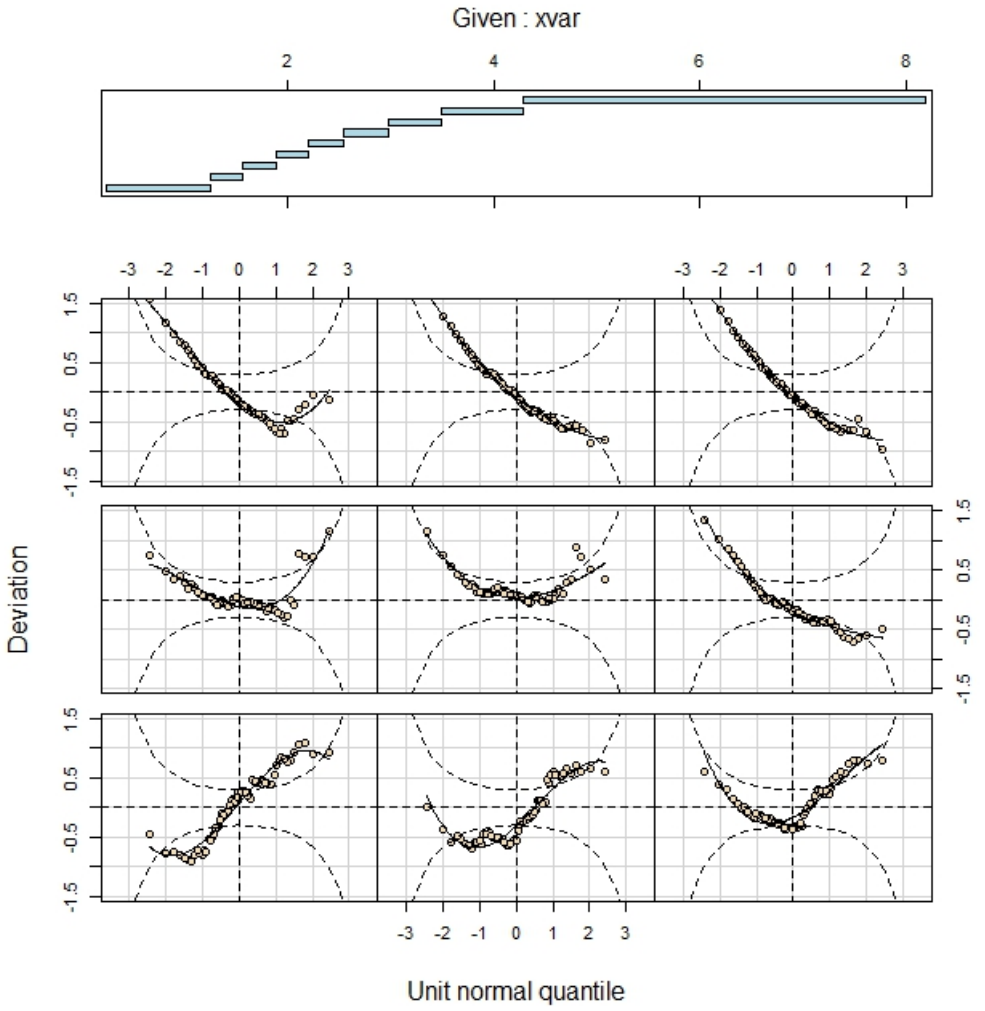

Figura 6.7: Worm plot por segmento do ajuste pelo polinômio fracionário para a média e linear para a dispersão.

Será ajustado um modelo de regressão paramétrica Weibull, sendo assim dadas pelas seguintes equações:

$$h\left(t|\mathbf{X}\right) = \gamma t^{\gamma-1} \exp\left[\left(\mathbf{X}\boldsymbol{\beta}\right)^{\gamma}\right] \tag{6.1}$$

$$S\left(t|\mathbf{X}\right) = \exp\left[-\left(\exp\left(\mathbf{X}\boldsymbol{\beta}\right)t\right)^{\gamma}\right]. \tag{6.2}$$

A análise tradicional, realizada por Carvalho et al (2011), [2], ocorre pelo comando do **R**:

$$survreg(formula = Surv(tempo, status) \sim idade + sexo + tratam, data = ipec, dist = "weibull")$$

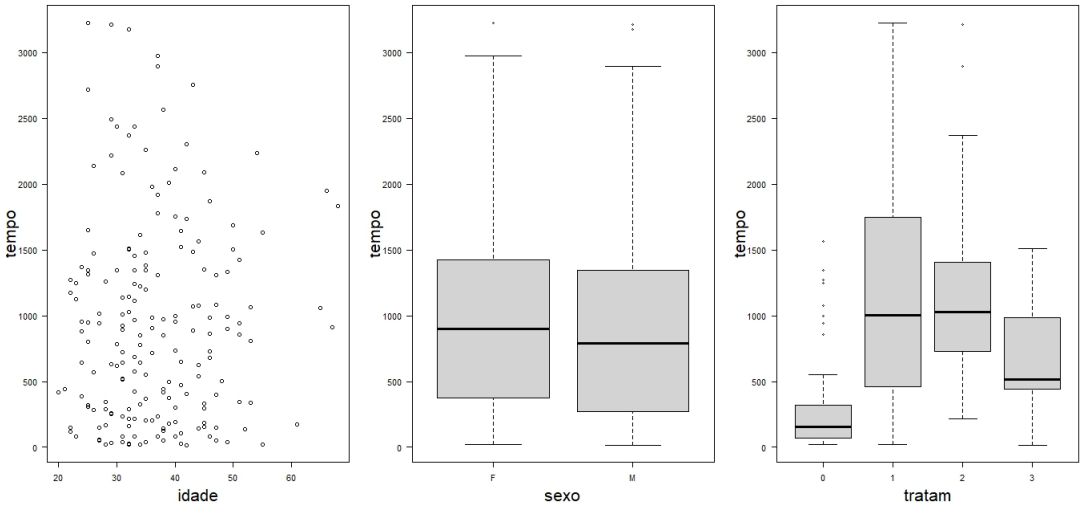

Figura 6.8: Relação entre as variáveis idade, sexo e tratam com a variável tempo

A análise por gamlss padronizar é feita pelo comando do **R**.

$$gamlss(Surv(tempo, status) \sim idade+sexo+tratam, sigma.fo =\sim 1, data = ipec, family = cens(WEI))$$

Na simulação com GAMLSS optou-se em manter o parâmetro de dispersão constante. A comparação entre as duas metodologias está na Tabela (6.4). A pequena diferença entre os valores se dá pela diferença entre os algoritmos em cada uma das funções. Pelo teste da Razão de

Tabela 6.4: Comparação entre as metodologias de análise de sobrevivência no R.

	survreg	gamlss
Intercept	5,97802	5,978713
idade	0,00903	0,008968
sexoM	-0,20476	-0,202474
tratam1	1,64172	1,639752
tratam2	2,83399	2,820017
tratam3	3,37902	3,308075
γ	0,88233	0,8772
$l(\boldsymbol{\theta})$	-741,4	-741,37

Verossimilhança temos ($pvalor$ = 0,98), portanto verifica-se que os dois modelos são equivalentes.

6.3 APÊNDICE 3: ESTADIAMENTO DA DOENÇA

Os estagio do câncer de melanoma é definido pelo sistema de classificação é o TNM, em que a classificação T fornece detalhes do tumor primário, N detalhes dos linfonodos regionais e M detalha a doença sistêmica (metástases). Para criar a variável estágio seguimos o trabalho de Gomes et al (2017) [12] que está de acordo com a AJCC (American Joint Comission on Cancer), que descreve a evolução clinica da doença. A classificação utilizada na presente investigação está resumida na Tabela (6.9).

ESTADIAMENTO CLÍNICO (cTNM) | 8ª EDIÇÃO AJCC – 2017			
ESTADIO	**T**	**N**	**M**
0	Tis	N0	M0
IA	T1a	N0	M0
IB	T1b ou T2a	N0	M0
IIA	T2b ou T3a	N0	M0
IIB	T3b ou T4a	N0	M0
IIC	T4b	N0	M0
III	Qualquer T	≥ N1	M0
IV	Qualquer T	Qualquer N	M1

Figura 6.9: Estadiamento Clínico 8ª EDIÇÃO AJCC – 2017 (Gomes et al (2017) [12])

BIBLIOGRAFIA

[1] Berkson, J., and Gage, R. P. Survival curve for cancer patients following treatment. *J. Am. Stat.Assoc.* **47** (1952), 501 – 502.

[2] Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S., and Shimakura, S. E. *Análise de Sobrevivência: Teoria e aplicações em saúde*. Fiocruz, Rio de Janeiro, 2011.

[3] Castro, M., Cancho, V. G., and Rodrigues, J. A hands-on approach for fitting long-term survival models under the gamlss framework. *Computer Methods and Programs in Biomedicine* **97** (2010), 168 – 177.

[4] Chambers, J. P., and Hastie, T. J. *Statistical Models in S*. UK, 1992.

[5] da Saúde, M. Instituto nacional do câncer. `https://www.inca.gov.br/estimativa/taxas-brutas/melanoma-maligno-da-pele`, 2022.

[6] de Melanoma, G. B. Informações gerais sobre o melanoma. `https://gbm.org.br/o-melanoma/`, 2022.

[7] de São Paulo, F. O. Banco de dados do rhc. `http://www.fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/`, 2022.

[8] Dunn, P. K., and Smyth, G. K. Randomised quantile residual. *J. Comput. Graph. Statist* **5** (1996), 236–244.

[9] Eiler, P. H. C., and Marx, B. D. Flexible smoothing with b-splines and penalties. *Statistical Science* **11** (1996), 89–121.

[10] Esteva, A., Kuprel, B., NOVOA, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542(7639)** (2017), 115 – 118.

[11] Gande, S. J., Munoz, A., Saez, M., and Alonso, A. Use of the beta-binomial distribution to model the eect of policy changes on appropriateness of hospital stays. *Appl. Statist.* **45** (1996), 371 – 382.

[12] Gomes, E., Landman, G., Belfort, F., and Schmerling, R. Estadiamento do melanoma pela ajcc. *Melanoma.* **76** (2017), 1–7.

[13] Gonzales, J. F. B. *Modelos de Sobrevivência com fração de cura via partição Bayesiana*. PhD thesis, Universidade Federal de São Carlos - São Carlos/SP., 2014.

[14] Harvey, A. C. Estimating regression models with multiplicative heteroscedasticity. *Econometrica* **41** (1976), 461–465.

[15] Hastie, T., and Tibshiarani, R. Generalized additive model. *Statical Science* **1** (1986), 297–318.

[16] James, B. R. *Probabilidade: um curso em nível intermediário*. IMPA, Rio de Janeiro, 2015.

[17] Kleinbaum, D. G., and Klein, M. *Survival Analysis: A Self-Learning Text*. Springer, New York, 2020.

[18] Magalhães, M. N. *Probabilidade e Variáveis Aleatótias.* Edusp., São Paulo, 2015.

[19] Maller, R. A., and Zhou, S. Estimating the proportion of immunes in a censored sample. *Biometrika* **79** (1992), 731–739.

[20] McCulloch, C. E. *Generalized linear and mixed model*. John Wiley and Sons, New York, 2001.

[21] Nelder, J. A., and Wedderburn, R. W. M. Generalized linear modelst. *Journal of the Royal Statistical Society* **135(3)** (1972), 370–384.

[22] Ramires, T. G., Nakamura, L. R., Righetto, A. J., Pescim, R. R., Mazucheli, J., and Cordeiro, G. M. A new semiparametric weibull cure rate model: fitting different behaviors within gamlss. *Journal of Applied Statistics* (2019).

[23] Rigby, R. A., and Stasinopoulos , D. M. A semi-parametric additive model for variance heterogeneity. *Statist. Comput* **6** (1996), 57 – 65.

[24] Rigby, R. A., and Stasinopoulos , D. M. Generalized additive models for location scale and shape. *Journal of Applied Statistics* **54** (2005), 507 – 554.

[25] Rigby, R. A., and Stasinopoulos , D. M. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software.* **23** (2007).

[26] Rigby, R. A., and Stasinopoulos , D. M. Automatic smoothing parameter selection in gamlss with an application to centile estimation. *Statistical Methods in Medical Research* **23(4)** (2014), 318 – 332.

[27] Rodrigues, J., Cancho, V. G., Castro, M., E.van Hoof, and Louzada-Neto, F. On the unification of long-term survival models. *Statistics and Probability Letters* **79** (2008), 753 – 759.

[28] Royston, P., and Altman, D. G. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics* **43** (1994), 429 – 467.

[29] Silversmit, G., Jegou, D., Vaes, E., E.van Hoof, Goetghebeur, E., and van Eycken, L. Cure of cancer for seven cancer sites in the flemish region. *International Journal of Cancer* **140** (2016), 1102 – 1110.

[30] Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and Bastiani, F. *Flexible Regression and Smoothing Using GAMLSS in R.* Chapman and Hall, London, 2020.

[31] van Buuren, S. Worm plot to diagnose fit in quantile regression. *Statistical Modelling* **7** (2007), 363 – 376.

[32] Wade, A. W., and Ades, A. E. Age-related reference ranges : Signifcance tests for models and confidence intervals for centiles. *Statistics in Medicine* **13** (1994), 2359 – 2367.