Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

MODELOS ADITIVOS GENERALIZADOS PARA POSIÇÃO, ESCALA E FORMA (GAMLSS) NA MODELAGEM DA ÁREA MIOCÁRDICA SOB RISCO DE NECROSE

Cássio de Alcântara

Uberlândia-MG
2018

Cássio de Alcântara

MODELOS ADITIVOS GENERALIZADOS PARA POSIÇÃO, ESCALA E FORMA (GAMLSS) NA MODELAGEM DA ÁREA MIOCÁRDICA SOB RISCO DE NECROSE

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Edmilson Rodrigues Pinto

Uberlândia-MG

2018

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Prof. Dr. Edmilson Rodrigues Pinto

_____

Prof. Dr. Janser Moura Pereira

_____

Prof. Dr. Leandro Alves Pereira

Uberlândia-MG

2018

Agradecimentos

À Tamara e Valentina, meus portos seguros.
Ao meu orientador Edmilson Rodrigues Pinto pela paciência, disponibilidade, versatilidade e dedicação ao conduzir de forma segura na produção deste trabalho.
Aos meus colegas de curso, pelo companheirismo e auxílio nesta jornada.
Aos professores: Lúcio, Janser, Leandro, Maria Imaculada, dentre outros, pelo conhecimento compartilhado, bem como ao apoio acadêmico no decorrer deste curso.
Aos meus amigos Matheus e Moacir pela parceria.
A Luiz Matheus, por acompanhar de perto todo o processo.
A Lucas, pelos conselhos e ensinamentos. Aos meus familiares e amigos pelo apoio constante.

Resumo

O presente trabalho tem como objetivos estudar a família dos Modelos Aditivos Generalizados para Posição, Escala e Forma (GAMLSS) e aplicar em um banco de dados, referente a pacientes que sofreram de Infarto Agudo do Miocárdio (IAM), onde a variável resposta é limitada ao intervalo contínuo $(0, 1)$. Desta forma, fez-se necessário o estudo de distribuições que se comportam bem com este tipo de varíavel, onde se destacam a Beta e a Beta Generalizada Tipo 1. Modelo de regressão beta foram obtidos para comparação dos ajustes. Os modelos GAMLSS resultantes modelaram dois parâmetros, μ e σ, e foi alcançado os pressupostos de normalidade, independência e homocedasticidade dos resíduos para ambos. Os valores de GAIC e R^2 generalizado foram superiores aos dos modelos de regressão beta criados, além de apresentarem melhores resultados na análise de resíduos. Conclui-se então que os modelos GAMLSS se apresentam como uma ferramenta poderosa no ajuste desses modelos, uma vez que auxilia o pesquisador em etapas da modelagem que muitas vezes são feitas de modo intuitivo além de modelar não só a média μ, como também a dispersão σ, a assimetria ν e a curtose τ, abrindo possibilidades de atingir bons ajustes sem a necessidade de excluir observações dos bancos de dados.

Palavras-chave: GAMLSS; infarto do miocárdio; regressão beta; beta generalizada tipo 1.

Abstract

The aim of this work is to study the Family of Generalized Additive Models for Position, Scale and Shape (GAMLSS) and apply it to a database, referring to patients who suffered from acute myocardial infarction (AMI), whose the response variable is limited to the continuous interval $(0, 1)$. In this way, it was necessary to study distributions that behave well with this type of variable, which stand out Beta and Generalized Beta Type 1. Beta regression models were obtained for comparison of the adjustments. The resulting GAMLSS models modeled two parameters, μ and σ, and the assumptions of normality, independency and homoscedasticity of the residues for both were reached. The generalized GAIC and R^2 values were superior to those of the beta regression models created, in addition to presenting better results in residue analysis. It is concluded that the GAMLSS models present themselves as a powerful tool in the adjustment of these models, since it assists the researcher in steps of the modeling that are often done in an intuitive way besides modeling not only the average μ, but also also the dispersion σ, asymmetry ν and kurtosis τ, opening possibilities to reach good adjustments without the need to exclude observations from the databases.

Keywords: GAMLSS; myocardial infarction; beta regression; generalized beta type 1.

Sumário

Lista de Figuras

Lista de Tabelas

1. INTRODUÇÃO

Por muitos anos, para descrever grande parte dos fenômenos aleatórios, foram utilizados os modelos normais lineares [9]. Quando o fenômeno em estudo não apresentava uma resposta para a qual a suposição de normalidade fosse razoável, tentava-se alguma transformação a fim de alcançar a normalidade procurada.

Com os avanços computacionais, alguns modelos que exigiam a utilização de esquemas iterativos para a estimação dos parâmetros começaram a ser mais empregados [6], como o modelo normal não-linear e os modelos não-lineares da família exponencial.

Dentre as técnicas de modelagem de regressão univariada, os modelos lineares generalizados (GLM) e os modelos aditivos generalizados (GAM) ocupam lugar de destaque na literatura. Em ambos os casos, assume-se que a distribuição da variável resposta pertença à família exponencial e a variância (σ), assimetria (ν) e curtose (τ) não são modeladas explicitamente a partir das variáveis explicativas, mas implicitamente a partir da relação com a média (μ).

Contudo, objetivando superar algumas das limitações associadas aos modelos acima descritos, Rigby e Stasinopoulos (2005) [13] introduziram os modelos aditivos generalizados para posição, escala e forma (GAMLSS). Nesta família, os parâmetros da distribuição condicional de Y podem ser modelados em função das variáveis explicativas através de um preditor linear η, que é composto por dois componentes: um paramétrico e outro não-paramétrico. Este último permite associar funções suavizadoras e a inclusão de termos de efeitos aleatórios.

Nos modelos GAMLSS a distribuição da variável resposta, Y, percence a uma família mais ampla do que a família exponencial, denominada família $\mathcal{D}$. A variável resposta Y tem distribuição $D(Y|\theta_i)$, $i = 1, ..., p$, em que $D \in \mathcal{D}$ pode ser qualquer distribuição. Além disso, a parte sistemática do modelo permite a modelagem de todos os parâmetros da distribuição condicional de Y.

De acordo com Rigby e Stasinopoulos (2005) [13], os modelos GAMLSS são adequados para a modelagem de variável resposta que não segue uma distribuição da família exponencial (por exemplo, leptocúrtica ou platicúrtica e/ou com assimetria positiva ou negativa) ou que exibem heterogeneidade, (por exemplo, quando a escala ou a forma da distribuição da variável resposta mudam com as variáveis explanatórias).

Este trabalho é motivado por um conjunto de dados composto por 64 pacientes com diagnóstico diferencial de Infarto Agudo do Miocárdio (IAM), que constitui-se de necrose miocárdica, proveniente de isquemia. Esta patologia é considerada como um dos principais problemas de saúde pública, causando milhares de mortes anuais pelo mundo. A estimativa precoce e correta

da área sob risco de necrose (ARN) no IAM possibilita ao médico a condução de tratamento adequado e eficiente ao paciente infartado.

O banco de dados utilizado já fora analisado utilizando o modelo de regressão beta e os resultados, publicados em [10]. Entretanto, na análise considerada, foram excluídos alguns registros do banco de dados, o qual permitiu um melhor ajuste, porém pode ter acarretado em perda de informação valiosa não considerada no modelo, o qual modelou apenas a média μ.

Desta forma, os objetivos deste trabalho são:

- Estudar a estrutura da família GAMLSS;
- Estudar médotos de estimação, distribuições, seleção de modelos e análise de resíduos da família GAMLSS;
- Criar modelos GAMLSS e de regressão beta para estimar a proporção da área miocárdica sob risco de necrose em pacientes que sofreram IAM sem a necessidade de exclusão de observações do banco de dados e comparar os seus resultados.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo iremos apresentar mais sobre os modelos GAMLSS, embasado na teoria de Rigby e Stasinopoulos (2005) [13].

2.1 MODELOS ADITIVOS GENERALIZADOS PARA POSIÇÃO, ESCALA E FORMA (GAMLSS)

2.1.1 DEFINIÇÃO

Nos modelos aditivos generalizados para posição, escala e forma (GAMLSS), termos paramétricos, aditivos e aleatórios são utilizados para modelar p parâmetros, $\boldsymbol{\theta}^\top = (\theta_1, ..., \theta_p)$, de uma função densidade de probabilidade $f(y|\boldsymbol{\theta})$, onde $\boldsymbol{y}^\top = (y_1, ..., y_n)$ é o vetor da variável resposta de interesse. Considerando as variáveis respostas y_i, $i = 1, ..., n$, independentes e condicionais a $\boldsymbol{\theta}^i$, isto é, $f(y_i|\boldsymbol{\theta}^i)$, em que $\boldsymbol{\theta}^{i\top} = (\theta_{i1}, ..., \theta_{ip})$ é o vetor de p parâmetros relacionado às variáveis explanatórias e termos aleatórios. Vale enfatizar que, quando os valores assumidos pelas variáveis são estocásticos, ou as observações y_i dependem de seus valores passados, então $f(y_i|\boldsymbol{\theta}^i)$ é interpretada como sendo condicional a estes valores.

Seja $\boldsymbol{y}^\top = (y_1, ..., y_n)$ o vetor de observações da variável resposta, para $k = 1, 2, ..., p$, considere uma função monótona $g_k(.)$ que relaciona o k-ésimo parâmetro $\boldsymbol{\theta}_k$ às variáveis explanatórias e efeitos aleatórios através de um modelo aditivo escrito da forma:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.1}$$

em que $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores $(n \times 1)$, $\boldsymbol{\beta}_k^\top = (\beta_{1k}, ..., \beta_{J_k'k})$ é um vetor de parâmetros de tamanho J_k', $\boldsymbol{X}_k$ e $\boldsymbol{Z}_{jk}$ são matrizes de planejamento (covariáveis) conhecidas e de ordens $(n \times J_k')$ e $(n \times q_{jk})$, respectivamente. Já $\boldsymbol{\gamma}_{jk}$ é uma variável aleatória q_{jk}-dimensional para o qual é usual assumir $\gamma_{jk} \sim N_{qjk}(0, \boldsymbol{G}_{jk}^{-1})$, onde $\boldsymbol{G}_{jk}^{-1}$ é a inversa (generalizada) de $\boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$, que pode depender de um vetor de hiperparâmetros $\boldsymbol{\lambda}_{jk}$. O modelo definido em (2.1) é denominado GAMLSS por [13].

Os vetores $\boldsymbol{\gamma}_{jk}$, $j = 1, ..., J_k$, podem ser manipulados e combinados em um único vetor $\boldsymbol{\gamma}_k$ e uma única matriz de covariáveis $\boldsymbol{Z}_k$. Entretanto, a formulação proposta em (2.1) é mais apropriada por dois motivos: facilita a utilização do algoritmo de auto reajuste (conhecido

como *backfitting*[1] para estimar os parâmetros e permite que combinações de diferentes termos aditivos e/ou efeitos aleatórios sejam facilmente incorporados ao modelo [13].

No caso em que $J_k = 0$, não há termos aditivos associados aos parâmetros da distribuição. Então, o modelo se reduz a um modelo linear completamente paramétrico dado por

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

Quanto $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, em que $\boldsymbol{I}_n$ é uma matriz identidade de ordem $n \times n$, e $\boldsymbol{\gamma}_{jk} = \boldsymbol{h}_{jk} = h_{jk}(x_{jk})$ para todas as combinações de j e k no modelo (2.1), temos

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk}), \tag{2.3}$$

em que x_{jk}, para $j = 1, ..., J_k$ e $k = 1, ..., p$, são vetores de tamanho n. A função h_{jk} é uma função desconhecida da variável explanatória x_{jk} e $\boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$, é um vetor que avalia h_{jk} em x_{jk}. Neste caso, assume-se que os vetores x_{jk} são conhecidos e o modelo (2.3) é denominado GAMLSS aditivo semi-paramétrico linear.

O modelo (2.3) pode ser estendido para permitir a inclusão de termos não-lineares na modelagem dos k parâmetros da distribuição, na forma

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(x_{jk}) \tag{2.4}$$

em que h_k, $k = 1, ..., p$ são funções não-lineares e $\boldsymbol{X}_k$ é uma matriz de covariáveis conhecida de ordem $n \times J_k''$. O modelo definido em (2.4) é designado de GAMLSS aditivo semiparamétrico não-linear. Se $J_k = 0$, o modelo (2.4) transforma-se num GAMLSS paramétrico não-linear, que é dado por:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) \tag{2.5}$$

Quando $h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) = \boldsymbol{X}_k^\top\boldsymbol{\beta}_k$, $k = 1, ..., p$, então o modelo definido em (2.5) é conhecido como modelo paramétrico linear (2.2). Note que alguns termos de $h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k)$ podem ser lineares, resultando em um modelo GAMLSS com a combinação de termos paramétricos lineares e não-lineares.

Na literatura é comum encontrar trabalhos que atribuem quatro parâmetros ($p = 4$), comumente caracterizados por posição (μ), escala (σ), assimetria (ν) e curtose (τ). Enquanto os dois primeiros parâmetros populacionais θ_1 e θ_2 no modelo (2.1), denotados por μ e σ, são referidos por parâmetros de posição (ou locação) e escala, respectivamente. No entando, os dois últimos $\nu = \theta_3$ e $\tau = \theta_4$ são denominados parâmetros de forma. Com isto, define-se

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \textstyle\sum_{j=1}^{J1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \textstyle\sum_{j=1}^{J2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \end{aligned}$$

[1] *backfitting* é um processo de ajuste iterativo que busca minimizar uma função de perda em relação à cada uma das funções até a convergência. Para mais detalhes, ver [7].)

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3},$$
$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}.$$

2.1.2 ESTIMAÇÃO

No caso do modelo GAMLSS paramétrico, mostrado na equação (2.2), a estimação é realizada através do método da máximo verossimilhança, sendo que, de acordo com [13], a função de máxima verossimilhança é dada por:

$$\ell = \sum_{i=1}^{n} log f(y_i|\theta^i), \tag{2.6}$$

em que f representa a função densidade de probabilidade da variável resposta.

Para modelos não paramétricos é necessário recorrer ao método da máxima verossimilhança penalizada, ℓ_p [13]:

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \lambda_{jk}\gamma_{jk}^{\top}\boldsymbol{G}_{jk}\gamma_{jk} \tag{2.7}$$

onde $\ell = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta^i})$ é a função de log-verossimilhança dos dados condicionais a $\boldsymbol{\theta^i}$, $i = 1, ..., n$.

Para a maximização da função de verossimilhança dada em (2.7), no R, podem ser utilizados dois algoritmos: CG e RS. O primeiro é uma generalização do algoritmo proposto por [3], este usa a primeira derivada e o valor esperado ou aproximado das derivadas de segunda ordem e das derivadas cruzadas da função de log-verossimilhança em relação aos parâmetros da distribuição $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$, para uma distribuição com quatro parâmetros. Entretanto, para muitas funções de densidade de probabilidade, $f(y|\boldsymbol{\theta})$, os parâmetros $\boldsymbol{\theta}$ são ortogonais, ou seja, os valores esperados das derivadas cruzadas da função de log-verossimilhança são iguais a 0. Neste caso é utilizado o algoritmo RS, que é uma generalização do algoritmo usado por [12] no ajuste da média e da dispesão de modelos aditivos e que, ao contrário do CG, não necessita das derivadas cruzadas da função de log-verossimilhança penalizada. Mais detalhes sobre os algoritmos CG e RS podem ser obtidos em [13].

Para ambos os algoritmos a estimação dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ é feita através da fixação do hiper-parâmetro $\boldsymbol{\lambda}$. No entanto, a estimação dos hiper-parâmetros $\boldsymbol{\lambda}$ pode ser feita de forma local (dentro dos algoritmos RS ou CG) ou de forma global (através da função **find.hyper()** do pacote gamlss), sendo que [13] recomendam a utilização dos metódos locais, uma vez que são mais rápidos e normalmente produzem resultados semelhantes aos métodos globais.

2.1.3 TERMOS ADITIVOS DO MODELO

Nos modelos GAMLSS, todos os parâmetros da distribuição podem ser modelados pelas covariáveis através de relações na forma linear e/ou não-linear e/ou através de funções suavizadoras não-paramétricas. Uma relação não-linear pode ser paramétrica não-linear ou um

suavizador. Estas relações afetam cada um dos valores preditos de cada parâmetro da distribuição, resultando na alteração da forma da distribuição da variável dependente [13].

Relação linear paramétrica

A relação linear considerada nos modelos GAMLSS é semelhante à dos modelos lineares generalizados. Quando não se verifica a linearidade da relação entre a variável resposta e determinada covariável é comum transformar esta última utilizando polinômios ou através de técnicas de suavização.

Relação não-linear paramétrica

Um exemplo desta relação são os polinômios aplicados às variáveis indepentendes que conferem certa flexibilidade à curva de regressão através da potência definida do polinômio. Existem diferentes tipos de polinômios, como os ortogonais, *fractional*, *piecewise* e *B-splines* [13].

Suavizadores

Os suavizadores não assumem a forma paramétrica que relaciona a variável resposta com as covariáveis, eles permitem que os dados determinem qual é essa relação funcional. Existem diversas funções suavizadoras disponíveis no pacote GAMLSS do R, estas são divididas em suavizadores penalizados (por exemplo, *cubic splines* e *tensor product splines*) e os restantes (por exemplo, *neural networks*). Os suavizadores penalizados utilizam a penalização quadrática para controlar a quantidade de suavização, e os restantes utilizam penalizações não quadráticas para obter afunção suavizadora.

Um detalhe importante sobre os suavizadores é que, quando são utilizados nos modelos GAMLSS, é preciso ter atenção à análise da saída do modelo obtido pelo programa R. Este decompõe o suavizador na sua parte 'linear' e parte 'não-linear', apresentando apenas o coeficiente e erro padrão da parte 'não-linear' [13].

Neste trabalho será abordado apenas o suavizador penalizado univariado denominado *cubic spline*, que é, de acordo com [14], um dos mais importantes do pacote GAMLSS do R devido à sua flexibilidade e à possibilidade de ser aplicado em diversas situações. Para mais detalhes sobre os demais suavizadores consultar [14].

A solução dos suavizadores penalizados univariados é o resultado da minimização da quantidade Q mostrada em (2.8), em relação a γ:

$$Q = (y - \boldsymbol{Z}\boldsymbol{\gamma})^\top \boldsymbol{W}(y - \boldsymbol{Z}\boldsymbol{\gamma}) + (\lambda \boldsymbol{\gamma}^\top \boldsymbol{G}\boldsymbol{\gamma}), \tag{2.8}$$

sendo $\boldsymbol{Z}$ a matriz de dimensão $n \times p$, já definida nos modelos GAMLSS em (2.1), $\boldsymbol{\gamma}$ o vetor de parâmetros de dimensão p a serem estimados, $\boldsymbol{W}$ a diagonal da matriz dos pesos de dimensão $n \times p$, λ o parâmetro suavizador e y a variável resposta.

A solução do problema de minimização de (2.8) é dada por:

$$\hat{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \lambda \boldsymbol{G})^{-1} \boldsymbol{Z}^\top \boldsymbol{W} y. \tag{2.9}$$

Diferentes $\boldsymbol{Z}$ e $\boldsymbol{G}$ produzem diferentes suavizadores e $\boldsymbol{W}$ é utilizado no algoritmo *backfitting* do modelo GAMLSS. Os valores ajustados obtidos são dados por [13]:

$$\hat{y} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \lambda \boldsymbol{G})^{-1} \boldsymbol{Z}^\top \boldsymbol{W} y = \boldsymbol{S} y, \tag{2.10}$$

onde $\boldsymbol{S}$ representa a matriz de suavização. O traço da matriz $\boldsymbol{S}$ é utilizado para determinar os graus de liberdade do suavizador [13], onde:

$$tr(\boldsymbol{S}) = tr[\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \lambda \boldsymbol{G})^{-1} \boldsymbol{Z}^\top \boldsymbol{W}]. \tag{2.11}$$

A estimação dos suavizadores, no caso do *cubic spline*, é feita através da penalização da segunda derivada da função de verossimilhança. Para mais detalhes, consultar [14].

2.1.4 DISTRIBUIÇÕES DISPONÍVEIS NO GAMLSS

A função densidade de probabilidade $f(\boldsymbol{y}|\boldsymbol{\theta})$ no modelo (2.1) pode pertencer a uma família de distribuições bastante geral sem que haja a obrigatoriedade de uma forma explícita para $\boldsymbol{y}$. No *software R*, a única restrição para a implementação dos modelos GAMLSS é que as primeiras derivadas de $f(\boldsymbol{y}|\boldsymbol{\theta})$, com relação aos parâmetros $(\boldsymbol{\theta})$, sejam calculáveis. Derivadas explícitas são preferíveis, mas é possível utilizar funções numéricas para o cálculo dessas derivadas.

De forma geral, os modelos GAMLSS atribuem à variável resposta distribuições de probabilidade que pertencem à família $\mathcal{D}$, a qual englobam distribuições da família exponencial, entre outras. Denotamos esta família de distribuições como:

$$\boldsymbol{y} \sim \mathcal{D}\{g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p\},$$

onde $\theta_1, ..., \theta_p$, são parâmetros de $\mathcal{D}$, $g_1, ..., g_p$ são funções de ligação e, $t_1, ..., t_p$ são fórmulas dos modelos para os termos explanatórios e efeitos aleatórios nos preditores $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_p$, respectivamente.

As Tabelas 2.1 e 2.2 exibem algumas famílias de distribuições contínuas e discretas, respectivamente, que se encontram implementadas no *software R*. Além disso, também é possível o ajuste de versões truncadas, censuradas ou de misturas finitas das distribuições.

Tabela 2.1: Exemplos de famílias de distribuições contínuas implementadas no pacote gamlss.

Distribuição	**Nomenclatura**	**Função de ligação de**			
		μ	σ	ν	τ
beta	BE()	logito	logito	-	-
beta inflacionada (em 0)	BEOI()	logito	log	logito	-
beta inflacionada (em 1)	BEZI()	logito	log	logito	-
beta inflacionada (em 0 e 1)	BEINF()	logito	logito	log	log
beta generalizada tipo 1	GB1()	logito	lgito	log	log
Box-Cox (Cole & Green)	BCCG()	identidade	log	identidade	-
Box-Cox exponencial potência	BCPE()	identidade	log	identidade	log
Box-Cox-t	BCT()	identidade	log	identidade	log
exponencial	EXP()	log	-	-	-
exponencial gaussiana	exGAUS()	identidade	log	log	-
gamma	GA()	log	log	-	-
gamma generalizada	GG()	log	log	identidade	-
gaussiana inversa	IG()	log	log	-	-
gaussiana inversa generalizada	GIG()	log	log	identidade	-
Gumbel	GU()	identidade	log	-	-
Gumbel reversa	RG()	identidade	log	-	-
Johnson's SU	JSU()	identidade	log	identidade	log
logística	LG()	identidade	log	-	-
log normal	LOGNO()	log	log	-	-
normal	NO()	identidade	log	-	-
shash	SHASH()	identidade	log	log	log
Weibull	WEI	log	log	-	-

Tabela 2.2: Exemplos de famílias de distribuições discretas implementadas no pacote gamlss.

Distribuição	**Nomenclatura**	**Função de ligação de**		
		μ	σ	ν
beta binomial	BB()	logito	log	-
binomial	BI()	logito	-	-
binomial negativa tipo 1	NBI()	log	log	-
binomial negativa tipo 2	NBII()	log	log	-
Delaporte	DEL()	log	log	logito
Poisson Inversa Gaussiana	PIG()	log	log	-
Poisson	PO()	log	-	-
Sichel	BCT()	log	log	identidade

2.1.5 SELEÇÃO DE MODELOS

Considere que $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ representa um modelo GAMLSS, onde $\mathcal{D}$ especifica a distribuição da variável resposta, $\mathcal{G}$ o conjunto das funções de ligação $(g_1, ..., g_p)$ para os parâmetros $(\theta_1, ..., \theta_p)$, $\mathcal{T}$ define o conjunto de termos preditores $(t_1, ..., t_p)$ para os preditores $(\eta_1, ..., \eta_p)$ e $\boldsymbol{\lambda}$ explicita o conjunto de hiperparâmetros.

O processo de construção de um modelo GAMLSS, em um determinado banco de dados, consiste em comparar diversos modelos concorrentes onde diferentes combinações dos componentes $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ foram utilizadas. Para tal, pode-se utilizar o critério de informação

de Akaíke generalizado (GAIC) [1], que é definido como $GAIC(a) = GD + a$, em que GD é o desvio global ajustado, $GD = -\ell(\widehat{\boldsymbol{\theta}})$, onde $\ell(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \ell(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}})$, em que a é a quantidade de graus de liberdade efetivos utilizada no modelo proposto. O critério de informação de Akaike (AIC) e o critério Bayseano de Schwarz (SBC) são casos especiais de GAIC, em que a penalidade adotada é $a = 2$ e $a = log(n)$, respectivamente [13]. Também pode-se utilizar o R^2 generalizado [8] definido por

$$R^2 = 1 - \left(\frac{L(0)}{L(\hat{\theta})}\right)^{(2/n)} \tag{2.12}$$

onde $L(0)$ e $L(\hat{\theta})$ denotam as funções de verossimilhança do modelo nulo, com apenas o termo constante, e do modelo ajustado, respectivamente. Considera-se um modelo adequado aquele que apresenta menor valor para o GAIC e/ou maior para o R^2 generalizado.

SELEÇÃO DA DISTRIBUIÇÃO ($\mathcal{D}$)

A seleção da distribuição da variável dependente é feita através do GAIC. Diferentes modelos GAMLSS com diferentes distribuições são ajustados e comparados e é selecionado aquele com menor valor de GAIC [13].

No *software* **R**, as funções **fitDist()** e **histDist()** auxiliam na escolha da distribuição da variável resposta. A primeira utiliza a função **gamlss()** para ajustar diferentes distribuições à variável dependente. Os argumentos da função **fitDist** são o vetor dos valores da variável dependente, o valor da penalização, a, do critério GAIC e o tipo de distribuições a ajustar (*'realline'*, *'realplus'* ou *'realAll'*).

A função **histDist** permite visualizar diferentes distribuições ajustadas à variável dependente. Ela obtém falores constantes para os parâmetros da distribuição, necessitando apenas a variável Y e a distribuição que deseja visualizar.

SELEÇÃO DAS FUNÇÕES DE LIGAÇÃO ($\mathcal{G}$)

As funções de ligação para cada parâmetro da distribuição é usualmente determinada pela distribuição escolhida para a variável resposta. No GAMLSS, cada distribuição já tem as respectivas funções de ligação selecionadas para cada parâmetro da distribuição, como pode ser visto nas Tabelas 2.1 e 2.2.

SELEÇÃO DOS TERMOS ADITIVOS ($\mathcal{T}$)

Os termos aditivos a serem inseridos no modelo para cada parâmetro da distribuição θ_k, $k = 1, 2, 3, 4$, podem ser lineares como suavizadores. Para a respectiva distribuição da variável resposta, a seleção dos termos aditivos tem de ser feita para todos os parâmetros da distribuição. Os termos adicionados influenciam os parâmetros da distribuição de forma diferente. A biblioteca gamlss do *software* **R** disponibiliza algumas funções para seleção dos termos aditivos,

neste trabalho abordaremos apenas a **stepGAICAll.A()**. Para mais detalhes sobre as demais, consultar [13].

A função **stepGAICAll.A()** seleciona covariáveis utilizando o critério GAIC utilizando a seguinte estratégia:

1. faz a seleção das covariáveis para o parâmetro μ utilizando o método *foward* considerando constantes σ, ν e τ,
2. realiza o procedimento do passo 1 para o parâmetro σ, considerando ν e τ constantes, mas com μ já ajustado com as covariáveis selecionadas no passo anterior,
3. repete o procedimento para ν e τ,
4. aplica a seleção *backward* ao parâmetro ν, mantendo as covariáveis selecionadas nos passos anteriores,
5. repete o passo anterior para σ e μ, mantendo as covariáveis selecionadas para os parâmetros dos passos anteriores.

Em todas as etapas, como critério de decisão, utiliza-se o critério GAIC. O modelo final irá conter uma sub-seleção das covariáveis para cada parâmetro da distribuição não necessariamente igual [13].

SELEÇÃO DOS HIPERPARÂMETROS ($\boldsymbol{\lambda}$)

Os hiperparâmetros podem ser estimados ou fixados. A forma tradicional de fixação do hiperparâmetro é feita fixando o número de graus de liberdade [7]. No entanto, é desejável estimá-lo. O pacote GAMLSS consegue fazer a estimação de λ automaticamente através dos métodos de estimação GCV (*Generalized cross validation*), GAIC e método de máxima verosimilhança. Os autores aconselham o método local devido à sua rapidez e também porque consegue obter resultados semelhantes ao método global [13].

2.1.6 ANÁLISE DE RESÍDUOS

Para a análise de resíduos é utilizado os resíduos dos quantis aleatórios normalizados, introduzido por [4], e definido por

$$r_i = \Phi^{-1}\{F(y_i; \widehat{\theta})\} \tag{2.13}$$

onde Φ representa a função de distribuição acumulada de uma normal padrão, $F(.)$ é a função de distribuição acumulada adequada aos dados e $\widehat{\theta}$ o vetor de parâmetros. Note que, um modelo adequado tem os resíduos r_i seguindo a distribuição normal padrão.

Utilizando a função **residuals()** do *software* **R** é possível obter o vetor dos resíduos do modelo ajustado. Para analisar a normalidade dos resíduos utilizam-se métodos gráficos como, por exemplo, gráficos de resíduos *versus* os valores ajustados (ou *versus* índice), gráfico densidade de Kernel ou o *Worm plot* [13].

WORM PLOT

Os gráficos *Worm plot*, proposto por [2], são úteis para identificar regiões em que o modelo não é bem ajustado aos dados. O eixo vertical do *Worm plot* retrata, para cada observação, a diferença entre a sua localização nas distribuições teórica e empírica. Os pontos, quando observados em conjunto, formam uma curva que se assemelha a uma minhoca. A forma do gráfico indica como os dados se distanciam da distribuição assumida e, quando tomadas em conjunto, sugerem modificações úteis no modelo, como pode ser visto na Tabela 2.3. Assim, se os pontos se encontrarem situados no interior da banda de confiança de 95% (entre as duas curvas elípticas), o ajuste do modelo é satisfatório.

Tabela 2.3: Interpretação de vários padrões do *worm-plot*.

Forma	**Momento**	**Se**	**Então**
Interceptar	Média	a minhoca passa acima da origem,	a média ajustada é muito pequena.
		a minhoca passa abaixo da origem.	a média ajustada é muito grande.
Inclinação	Variância	a minhoca tem uma inclinação positiva,	a variância ajustada é muito pequena.
		a minhoca tem uma inclinação negativa,	a variância ajustada é muito grande.
Parábola	Assimetria	a minhoca tem formato de U,	a distribuição ajustada é assimétrica à esquerda.
		a minhoca tem formato de U invertido,	a distribuição ajustada é assimétrica à direita.
Curva S	Curtose	a minhoca tem uma forma em S à esquerda curvada para baixo,	as caudas da distribuição ajustada são muito leves.
		a minhoca tem uma forma em S à esquerda curvada para cima,	as caudas da distribuição ajustada são muito pesadas.

3. Metodologia

Neste capítulo será descrito o banco de dados trabalhado, mostrando as características das variáveis a serem estudadas. Posteriormente, serão apresentadas a distribuição Beta, assim como seu modelo de regressão, e a Beta Generalizada Tipo 1, pertencentes à família GAMLSS.

3.1 O banco de dados

Os dados recolhidos e utilizados para este estudo já foram previamente analisados e os resultados, publicados em [10].

O conjunto de dados é composto por 64 pacientes dinamarqueses com diagnóstico diferencial de infarto agudo no miocárdio (IAM). Estes foram divididos em dois grupos, de acordo com o tipo de infarto sofrido: o primeiro, com 36 pacientes que sofreram infarto inferior, e o segundo, com 28 pacientes que sofreram infarto anterior. Em ambos os grupos, a variável dependente y representa a porcentagem da área miocárdica sob risco de necrose (ARN).

As covariáveis são as derivações do eletrocardiograma, que, para o primeiro grupo, são as derivações D2, D3 e aVF e para o segundo grupo as derivações V1, V2, V3, V4, V5 e V6, as quais representam as derivações precordiais, e são relacionadas ao infarto de parede anterior, enquanto que D2, D3 e aVF representam derivações relacionadas à parede inferior do coração. Em ambos os casos, também foram consideradas as covariáveis idade (em anos) e sexo (onde 0 é masculino) do paciente. Para mais detalhes sobre o banco de dados, ver [10]. Os dados estão dispostos nas Tabelas B.1 e B.2.

3.2 *Software*

A análise dos dados foi implementada no programa estatístico R [15], versão 3.5.1. As bibliotecas utilizadas foram: **gamlss**, **lmtest** e **nortest**. No Apêndice A encontra-se o *script* de toda a análise efetuada neste estudo.

3.3 Distribuição Beta

A distribuição beta é muito flexível em situações onde a varíavel dependente Y é contínua e restrita ao intervalo $(0, 1)$, pois sua função de densidade pode assumir diferentes formas dependendo dos valores dos parâmetros que a compõem.

A variável Y segue uma distribuição beta com parâmetros p e q se sua função densidade de probabilidade é dada por:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{(p-1)}(1-y)^{(q-1)}, \tag{3.1}$$

onde $0 < y < 1$, $p, q > 0$ e $\Gamma(p)$ é a função gama no ponto p, que é dada por

$$\Gamma(p) = \int_0^\infty y^{(p-1)} e^{-y} dy. \tag{3.2}$$

A média e a variância de Y são dadas por:

$$E(Y) = \frac{p}{p+q}, \tag{3.3}$$

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{3.4}$$

Uma reparametrização que possibilita a modelagem da média da variável resposta através de uma estrutura de regressão e que envolve também um parâmetro de precisão foi proposta por Ferrari e Cribari Neto (2004) em [5].

Sejam $\mu = \frac{p}{(p+q)}$ e $\phi = p + q$, assim $p = \mu\phi$ e $q = (1-\mu)\phi$, logo as equações (3.3) e (3.4) são dadas por:

$$E(Y) = \mu, \tag{3.5}$$

$$Var(Y) = \frac{\mu(1-\mu)}{1+\phi} = \frac{1}{1+\phi} V(\mu) = \sigma^2 V(\mu), \tag{3.6}$$

onde $V(\mu) = \mu(1-\mu)$, $\sigma^2 = \frac{1}{1+\phi}$, μ é o parâmetro de posição, σ^2 é o parâmetro de dispersão e ϕ pode ser interpretado como o parâmetro de precisão. Neste caso, a função de densidade para Y apresenta a seguinte forma:

$$f(y;\mu,\sigma) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{3.7}$$

onde $0 < y < 1$, $0 < \mu < 1$ e $\phi > 0$.

Na família GAMLSS a distribuição beta assume que $\sigma^2 = \frac{1}{(1+\phi)}$ e $\phi = \frac{(1-\sigma^2)}{\sigma^2}$, sendo $\phi > 1$ e $0 < \sigma^2 < 1$. Assim, a função (3.1) pode ser reescrita como:

$$f(y;\mu,\sigma) = \frac{\Gamma(\frac{1-\sigma^2}{\sigma^2})}{\Gamma(\mu\frac{1-\sigma^2}{\sigma^2})\Gamma((1-\mu)\frac{1-\sigma^2}{\sigma^2})} y^{\mu(\frac{1-\sigma^2}{\sigma^2})-1}(1-y)^{(1-\mu)(\frac{1-\sigma^2}{\sigma^2})-1}, \tag{3.8}$$

onde $0 < y < 1$, $0 < \mu < 1$ e $0 < \sigma^2 < 1$.

Vale ressaltar que, na equação (3.8), caso ϕ seja um valor constante, o modelo de regressão coincide com o modelo de regressão beta proposto por [5]. Desta forma, neste trabalho será utilizado o GAMLSS para criar modelos de regressão beta ao invés de utilizar as funções da biblioteca **betareg**.

3.4 DISTRIBUIÇÃO BETA GENERALIZADA TIPO 1

A família da distribuição beta generalizada tipo 1, tem suporte $0 < Y < 1$, GB1$(\mu, \sigma^2, \nu, \tau)$, e define uma variável Z pertencente a uma distribuição Beta, BE(μ, σ^2) [11]:

$$Z = \frac{Y^\tau}{\nu + (1-\nu)Y^\tau}, \tag{3.9}$$

onde $0 < \mu < 1$, $0 < \sigma^2 < 1$, $\nu > 0$ e $\tau > 0$, sendo μ, σ^2, ν e τ os parâmetros de posição, dispersão, assimetria e curtose, respectivamente [11]. A função densidade de probabilidade de GB1 é dada por:

$$f(y; \mu, \sigma^2, \nu, \tau) = \frac{\tau \nu^q y^{\tau p - 1}(1 - y^\tau)^{q-1}}{B(p,q)[\nu + (1-\nu)y^\tau]^{p+q}}, \tag{3.10}$$

onde $p = \frac{\mu(1-\sigma^2)}{\sigma^2}$ e $q = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}$, $p > 0$, $q > 0$ e $B(p,q)$ é a função beta, que é dada por

$$B(p,q) = \int_0^1 t^{(p-1)}(1-t)^{q-1} dt. \tag{3.11}$$

Os parâmetros μ e σ^2 são adaptados para $\mu = \frac{\alpha}{(p+q)}$ e $\sigma^2 = \frac{1}{(p+q+1)}$. A distribuição beta é um caso especial da beta generalizada tipo 1 em que $\nu = 1$ e $\tau = 1$ [11].

4. RESULTADOS

Os resultados serão dispostos da seguinte maneira: primeiramente é realizada uma análise exploratória no conjunto de dados, seguido do ajuste de modelos de regressão beta e, posteriormente, GAMLSS.

4.0.1 ANÁLISE EXPLORATÓRIA

VARIÁVEL DEPENDENTE

As medidas descritivas da variável dependente, área sob risco de necrose, estão dispostas na Tabela 4.1.

Tabela 4.1: Medidas descritivas da área sob risco de necrose.

Infarto	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Inferior	0,0800	0,1600	0,2200	0,2758	0,3600	0,5400
Anterior	0,2100	0,3650	0,4150	0,4104	0,4600	0,6000

Para esta variável ainda foram construídos, para cada tipo de infarto, um histograma e um box-plot (Figura 4.1). Notamos que, em ambos os casos, a variável está limitada no intervalo $(0, 1)$, sendo que, para o infarto inferior, por apresentar média maior do que a mediana, apresenta assimetria à direita, o que nos dá indicativa de quais distribuições podem ser utilizadas na modelagem. Para tal, pode ser utilizado a beta, para modelar μ e σ^2, a beta generalizada tipo 1, para modelar μ, σ^2, ν e τ, ou pode ser feito truncamento de outras distribuições, como a Weibul ou a log normal, mas neste trabalho não será abordado truncamento de funções. Também nota-se que não há presença de valores extremos.

VARIÁVEIS INDEPENDENTES

As medidas descritivas das variáveis indpendentes contínuas de ambos os tipos de infarto estão dispostas na Tabela 4.2. Podemos observar, através da relação entre a média e a mediana, que as variáveis D2, D3, aVF, V2, V3, V4, V5 e V6 apresentam assimetria à direita (média maior do que a mediana), as variáveis Idade do infarto inferior e $V1$ apresentam assimetria à esquerda (média menor do que a mediana) e a variável Idade do infarto anterior apresenta simetria (média igual a mediana). Além disso, nota-se grande concentração de observações com 0 nas variáveis V5 e V6, fato que pode ser confirmado por ambas apresentarem 0 no 1^o quartil.

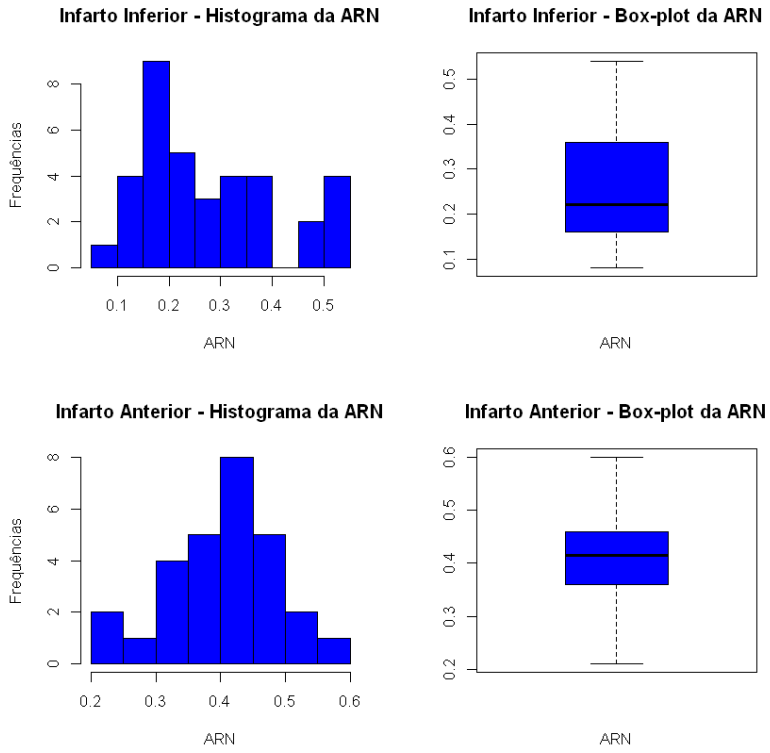

Figura 4.1: Histograma (esquerda) e *Box-plot* (direita) da variável ARN.

Tabela 4.2: Medidas descritivas das variáveis independentes contínuas.

Infarto	Variável	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Inferior	$Idade$	$39,0$	$52,0$	$62,0$	$61,0$	$70,0$	$79,0$
	$D2$	$0,5$	$1,8$	$2,5$	$2,6$	$3,1$	$6,5$
	$D3$	$0,5$	$2,0$	$3,0$	$3,2$	$4,6$	$9,0$
	aVF	$1,0$	$1,5$	$2,5$	$2,8$	$4,0$	$7,5$
Anterior	$Idade$	$46,0$	$55,0$	$63,0$	$63,0$	$69,0$	$89,0$
	$V1$	$0,0$	$1,0$	$1,5$	$1,4$	$2,0$	$3,0$
	$V2$	$0,0$	$2,0$	$3,7$	$3,8$	$5,0$	$8,0$
	$V3$	$1,0$	$2,5$	$3,2$	$4,1$	$5,1$	$13,0$
	$V4$	$0,0$	$1,9$	$2,2$	$3,1$	$5,0$	$8,0$
	$V5$	$0,0$	$0,0$	$1,0$	$1,4$	$2,1$	$4,5$
	$V6$	$0,0$	$0,0$	$0,0$	$0,4$	$1,0$	$2,0$

Em relação à variável dicotômica sexo é possível observar pela Tabela 4.3 que os indivíduos da amostra não se distribuem de forma equitativa entre as categorias desta variável, onde notamos uma presença maior de homens em ambos os tipos de infarto.

Tabela 4.3: Proporções de indivíduos para as categorias da variável sexo.

Infarto	Gênero	Total
Inferior	Masculino	23 (63, 89%)
	Feminino	13 (36, 11%)
Anterior	Masculino	23 (82, 14%)
	Feminino	5 (17, 86%)

No *box-plot* (Figura 4.2) é possível observar assimetria à esquerda em ambos os gráficos para a categoria do gênero feminino e do masculino. Além disso, para o gênero feminino ainda há a presença de valores extremos. Já no *box-plot* (Figura 4.3) a assimetria acontece apenas no gênero feminino e a presença de valores extremos no gênero masculino.

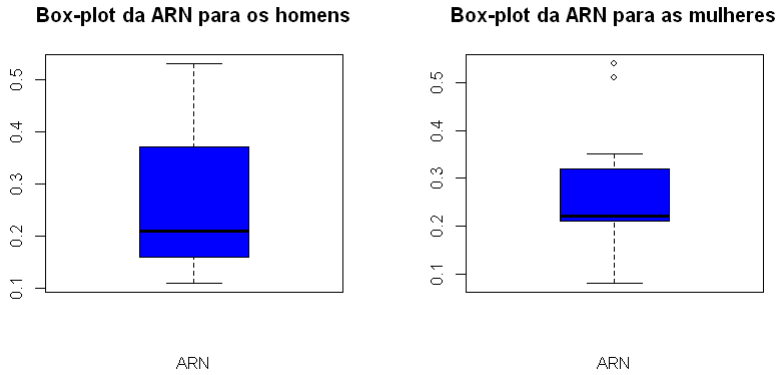

Figura 4.2: Infarto Inferior - *Box-plot* da área sob risco de necrose (ARN) para homens (esquerda) e mulheres (direita).

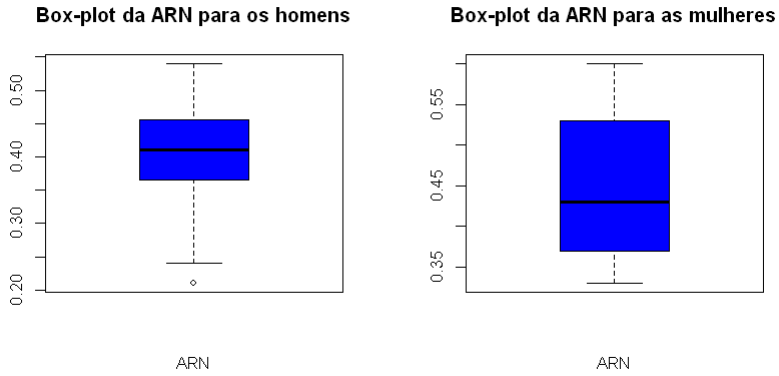

Figura 4.3: Infarto Anterior - *Box-plot* da área sob risco de necrose (ARN) para homens (esquerda) e mulheres (direita).

RELAÇÃO ENTRE AS VARIÁVEIS

Para verificar se as variáveis contínuas possuem relação linear com a variável resposta e entre si, foi construída uma matriz de correlação para cada tipo de infarto (Tabelas 4.4 e 4.5). Nestas,

verica-se que nenhuma das variáveis independentes apresentaram correlações superiores a $0,50$ com a variável dependente. Além disso, para o infarto inferior, observamos que as variáveis D2, D3 e aVF possuem correlação superior a $0,80$ entre si, o que indica a existência de colinearidade, portanto, na modelagem deste tipo de infarto, será utilizado apenas as covariáveis idade, sexo e D2. O mesmo ocorre para os pares de variáveis V4 com V5 e V2 com V3 do infarto anterior, onde não serão utilizadas as variáveis V2 e V5.

Tabela 4.4: Matriz de correlação das variáveis contínuas dos pacientes com infarto inferior.

	ARN	**idade**	**D2**	**D3**	**aVF**
ARN	$1,00$	–	–	–	–
idade	$0,22$	$1,00$	–	–	–
D2	$0,26$	$0,16$	$1,00$	–	–
D3	$0,18$	$0,17$	$0,91$	$1,00$	–
aVF	$0,21$	$0,25$	$0,87$	$0,89$	$1,00$

Tabela 4.5: Matriz de correlação das variáveis contínuas dos pacientes com infarto anterior.

	ARN	**idade**	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**
ARN	$1,00$	–	–	–	–	–	–	–
idade	$0,20$	$1,00$	–	–	–	–	–	–
V1	$-0,28$	$0,31$	$1,00$	–	–	–	–	–
V2	-0.42	$0,16$	$0,56$	$1,00$	–	–	–	–
V3	$-0,29$	$0,23$	$0,40$	$0,80$	$1,00$	–	–	–
V4	$-0,12$	$0,42$	$0,30$	$0,51$	$0,72$	$1,00$	–	–
V5	$-0,00$	$0,29$	$0,21$	$0,28$	$0,47$	$0,82$	$1,00$	–
V6	$0,31$	$-0,13$	$-0,24$	$-0,37$	$-0,25$	$0,10$	$0,48$	$1,00$

4.0.2 ANÁLISE DOS DADOS ATRAVÉS DO MODELO DE REGRESSÃO BETA

Para obter um modelo de regressão beta utilizando a biblioteca **gamlss** basta recorrer ao GAMLSS linear completamente paramétrico (2.2). Para tal, no processo de modelagem, não se utiliza de termos aditivos. Além disso, modela-se apenas a média μ.

ANÁLISE DO MODELO

A construção do modelo é efetuada com base nas variáveis de interesse, idade, sexo e D2 do infarto inferior, idade, sexo V1, V3, V4 e V6 do infarto anterior, para o parâmetro μ em cada banco de dados. Nas Tabelas 4.6 e 4.7 tem-se os modelos com todas as variáveis, para o parâmetro μ, seguidos dos modelos após aplicação da função **stepGAICALL.A**, assim como seus valores de GAIC e R^2 generalizado para comparação dos mesmos, para os infartos inferior e anterior, respectivamente. Podemos observar que, para ambos os infartos, o modelo

2 apresentou menor valor de GAIC e o R^2 generalizado não teve grande variação, sendo então selecionado como melhor modelo.

Tabela 4.6: Modelos de regressão beta para o infarto inferior.

Modelo	**Forma funcional**	**GAIC**	R^2 **gen.**	**Considerações**
1	$idade + sexo + D2$	$-43,4410$	$0,1239$	Modelo considerando todas as variáveis.
2	$D2$	$-46,1326$	$0,0914$	Modelo após aplicação da função **stepGAICALL.A**.

Tabela 4.7: Modelos de regressão beta para o infarto anterior.

Modelo	**Forma funcional**	**GAIC**	R^2 **gen.**	**Considerações**
1	$idade+sexo+V1+V3+V4+V6$	$-51,1512$	$0,2971$	Modelo considerando todas as variáveis.
2	$idade + V1 + V6$	$-55,2510$	$0,2477$	Modelo após aplicação da função **stepGAICALL.A**.

Tendo selecionado os modelos que apresentaram menor valor de GAIC para ambos os infartos, a expressão dos modelos de regressão beta foram, para o infarto inferior:

$$g_1(\mu) = \beta_{10} + \beta_{13}D2, \tag{4.1}$$

para o infarto anterior:

$$g_1(\mu) = \beta_{10} + \beta_{11}idade + \beta_{12}V1 + \beta_{15}V6, \tag{4.2}$$

Nas Tabelas 4.8 e 4.9 temos os valores das estimativas, erro padrão e valor p, onde podemos verificar que, para o modelo do infarto inferior, o valor p da variável D2 é superior a $0,05$ (Tabela 4.8), indicando que não é rejeitada a possibilidade de o coeficiente ser igual a zero, portanto, pode ser considerada como pouco relevante para este modelo de predição. O mesmo acontece para todas as variáveis no modelo do infarto anterior (Tabela 4.9)

Tabela 4.8: Estimativas dos parâmetros para $g(\mu)$ para o modelo beta do infarto inferior.

Variável	**Estimativa**	**Erro Padrão**	**valor p**
Intercepto	$-1,2807$	$0,1973$	$0,0000$
D2	$0,1228$	$0,0638$	$0,0629$

Tabela 4.9: Estimativas dos parâmetros para $g(\mu)$ para o modelo beta do infarto anterior.

Variável	Estimativa	Erro Padrão	valor p
Intercepto	$-0,9039$	$0,3392$	$0,0138$
idade	$0,0104$	$0,0053$	$0,0653$
V1	$-0,1278$	$0,0698$	$0,0804$
V6	$0,1438$	$0,0891$	$0,1201$

ANÁLISE DE RESÍDUOS

Para verificar se os resíduos seguem distribuição normal foram construídos *box-plot* e *QQ-plot*. Na Figura 4.4 estão representados os gráficos para análise dos resíduos obtidos pelo modelo do infarto inferior. No gráfico *QQ-plot* (Figura 4.4) é possível observar que a maioria dos resíduos estão próximos da reta diagonal ($y = x$), mas alguns estão bem distantes, além disso, no gráfico *box-plot*, os resíduos aparentam seguir uma distribuição assimétrica à direita além de nenhum ponto aparecer como extremo. Estes gráficos indicam que a distribuição dos resíduos possa não seguir distribuição normal.

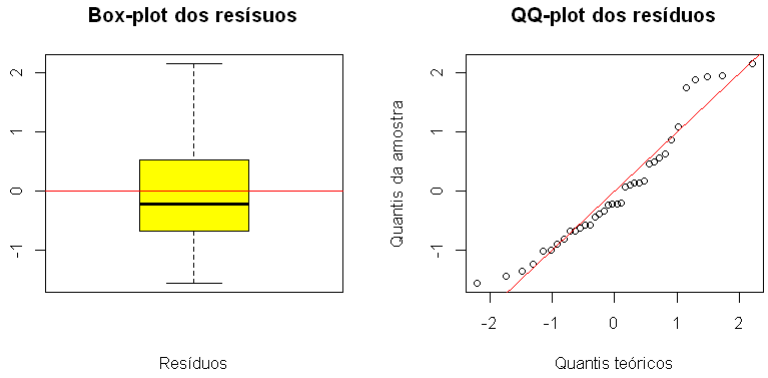

Figura 4.4: Box-plot (esquerda) e *QQ-plot* (direita) dos resíduos do modelo beta para o infarto inferior.

Na Figura 4.5 estão representados os gráficos para análise dos resíduos obtidos pelo modelo do infarto anterior. No gráfico *QQ-plot* (Figura 4.5) é possível observar que a maioria dos resíduos estão próximos da reta diagonal ($y = x$), mas alguns estão bem distantes, no gráfico *box-plot*, não observamos a presença de valores extremos. Estes gráficos indicam que a distribuição dos resíduos possa seguir distribuição normal.

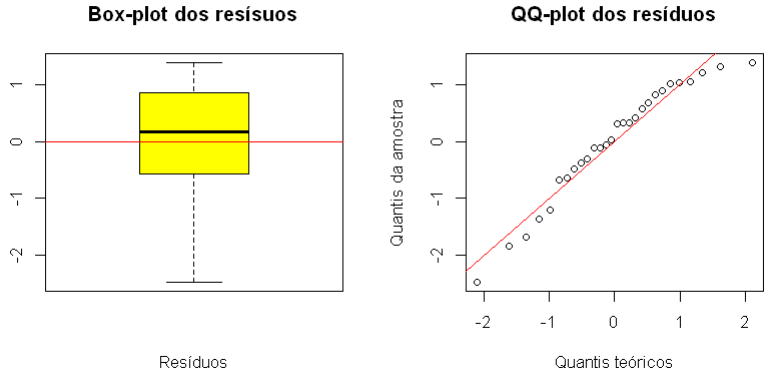

Figura 4.5: Box-plot (esquerda) e *QQ-plot* (direita) dos resíduos do modelo beta para o infarto anterior.

Através da função **plot()** é possível obter medidas descritivas dos resíduos. Estas estão dispostas na Tabela 4.10, onde podemos ver que os resíduos têm uma média de $-0,0050$ e variância de $1,0258$ para o modelo do infarto inferior e uma média de $0,0000$ e variância de $1,0369$ para o modelo do infarto anterior, o que sugere que os resíduos para os dois modelos seguem distribuição normal padrão.

Tabela 4.10: Medidas descritivas dos resíduos do modelo beta ajustado.

Infarto	**Medida**	**Valor**
Inferior	Média	$-0,0050$
	Variância	$1,0258$
	coef. de assimetria	$0,6334$
	coef. de curtose	$2,4864$
	coef. de correlação de Filliben	$0,9699$
Anterior	Média	$0,0000$
	Variância	$1,0369$
	coef. de assimetria	$-0,6485$
	coef. de curtose	$2,5300$
	coef. de correlação de Filliben	$0,9740$

A função **plot()** ainda mostra quatro gráficos: resíduos *versus* valores ajustados (*'Against Fitted Values'*), resíduos *versus index* (*'Against Index'*), gráfico de estimação não-paramétrica (suavizador de Kernel) da densidade dos resíduos (*'Density Estimate'*) e *'Normal Q-Q Plot'*. Sendo que este último já foi referido e analisado (Figuras 4.4 e 4.5 - direita).

Nas Figuras 4.6 e 4.7 estão apresentados os gráficos obtidos pela função **plot()**. Observando os gráficos do canto superior direito, na Figura 4.6, os valores dos resíduos quantílicos aparentam ter tendência decrescente, isto acontece na Figura 4.7. Nos gráficos do canto superior esquerdo não existe qualquer padrão para os resíduos. No gráfico localizado no canto inferior esquerdo das figuras, podemos observar que ambas as distribuições não apresentam forma muito semelhante à da função densidade da normal padrão. A homocedasticidade, independência e a normalidade

dos resíduos foram testadas pelos testes F, Durbin-Watson e Shapiro-Wilk, respectivamente, ao nível de 5% de significância, onde a hipótese de que os resíduos são homocedásticos foi rejeitada para o infarto inferior e não foi para o anterior, a independência não foi rejeitada para ambos os tipos de infarto e a homocedasticidade foi rejeitada para ambos os tipos de infarto, conforme pode ser visto na Tabela 4.11.

Tabela 4.11: Testes para normalidade e homocedasticidade dos resíduos do modelo beta ajustado.

Infarto	**Teste**	**Hipótese nula**	**valor p**
Inferior	Shapiro-Wilk para normalidade	Os resíduos seguem distribuição normal	$0,0332$
	Teste F para homoscedasticidade	Os resíduos são homocedásticos	$0,0157$
	Durbin-Watson para independência	Os resíduos são independentes	$0,9804$
Anterior	Shapiro-Wilk para normalidade	Os resíduos seguem distribuição normal	$0,1371$
	Teste F para homoscedasticidade	Os resíduos são homocedásticos	$0,0323$
	Durbin-Watson para independência	Os resíduos são independentes	$0,1481$

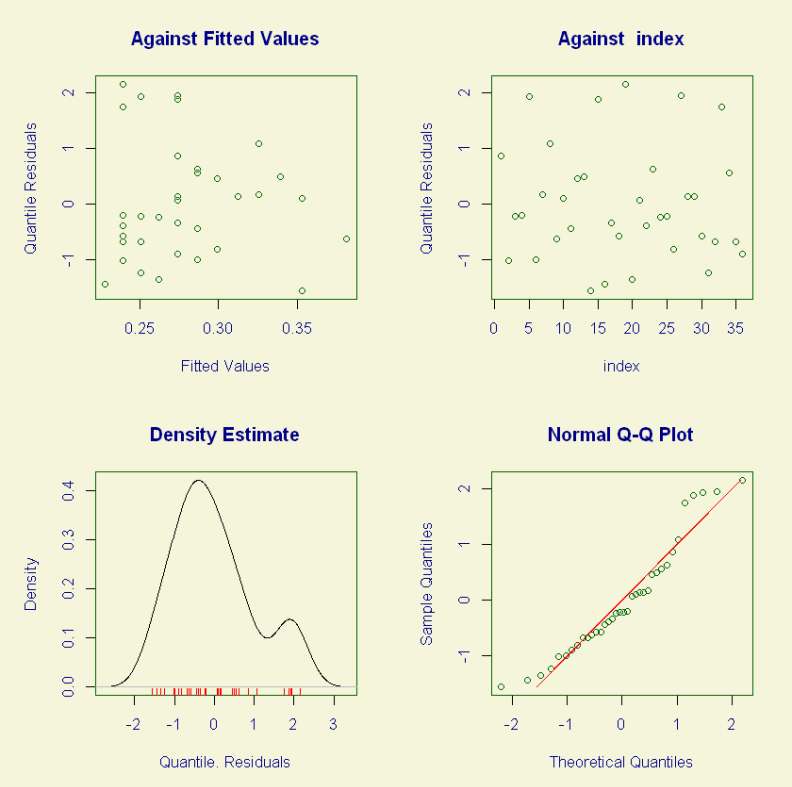

Figura 4.6: Gráficos dos resíduos obtidos através da função **plot()** do modelo beta para o infarto inferior.

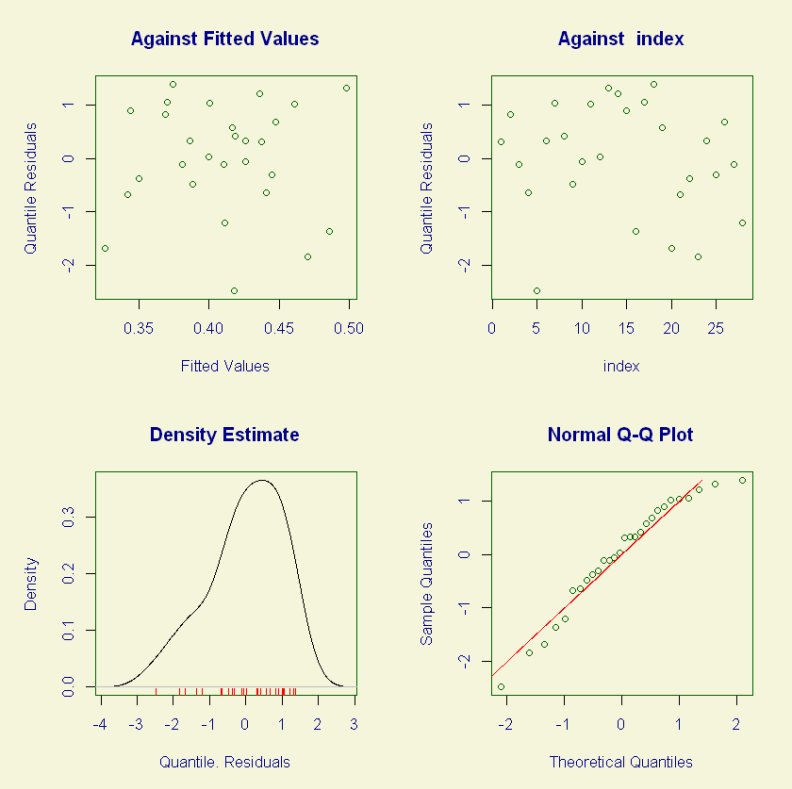

Figura 4.7: Gráficos dos resíduos obtidos através da função **plot()** do modelo beta para o infarto anterior.

Na Figura 4.8 estão representados os gráficos *worm plot* de cada modelo. No *worm plot* do modelo do infarto inferior, no lado esquerdo da figura, podemos observar que quase todos os pontos estão dentro da banda de confiança de 95%, além disso, os pontos estão em sua maioria concentrados próximos à reta $y = 0$, o que indica um ajuste razoável do modelo. No *worm plot* do modelo do infarto anterior, no lado direito da figura, podemos observar que todos os pontos estão dentro da banda de confiança de 95% e a maioria dos pontos estão concentrados próximos à reta $y = 0$, indicando um bom ajuste do modelo.

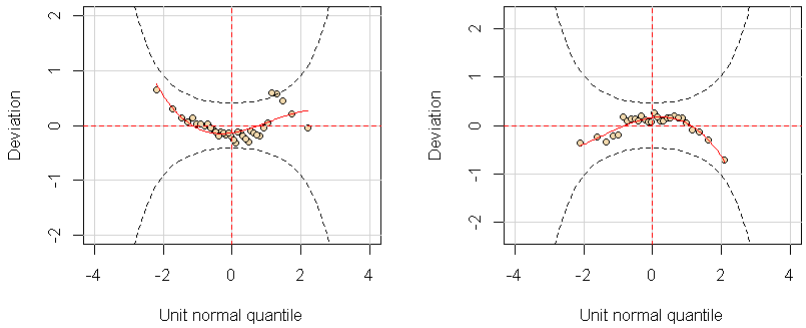

Figura 4.8: Gráficos *worm plot* dos modelos beta para o infarto inferior (esquerda) e anterior (direita).

4.0.3 ANÁLISE DOS DADOS ATRAVÉS DO GAMLSS

ESCOLHA DA DISTRIBUIÇÃO DA VARIÁVEL DEPENDENTE

Inicialmente foi realizada, para ambos os tipos de infarto, a estimação da distribuição da variável dependente área sob risco de necrose. A função **fitDist()** possui o argumento *type*, que teve necessidade de ser especificado para ajustar as funções contínuas com suporte $(0, 1)$ à variável dependente, selencionando a opção *real0to1*. Neste caso a função considera todas as funções que atendem esta condição, são elas: BE, BEOI, BEZI, BEINF0, BEINF1, GB1 e BEINF. Mas serão considerados apenas BE e GB1, pois a distribuição da variável resposta não apresenta comportamento que justifica considerar as betas infladas, uma vez que não há concentração em 0 ou em 1 (Figura 4.1).

Na Figura 4.9 temos o resultado da função **histDist()** para as distribuições *BE - Beta* e *GB1 - Beta Generalizada Tipo 1*, em ambas, a linha vermelha representa a densidade paramétrica e a azul a densidade estimada não-parametricamente. Ao observar a figura, notamos que a estimação pela distribuição beta aparementamente se ajusta melhor.

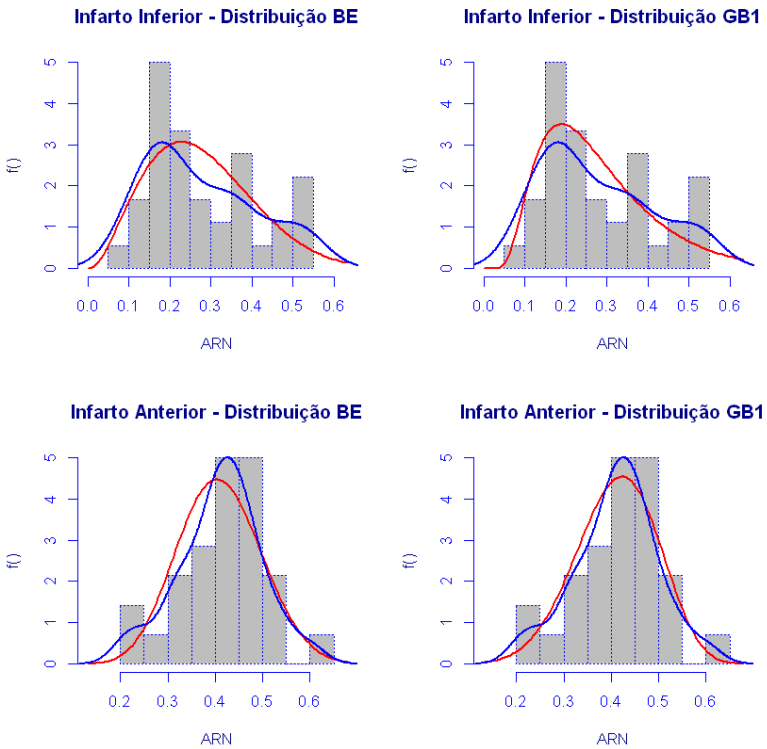

Figura 4.9: Histograma com a distribuição Beta ajustada (esquerda) e Histograma com a distribuição Beta Generalizada Tipo 1 ajustada (direita) da variável dependente para os dois tipos de infarto.

Todas as distribuições testadas conseguiram ser ajustadas à variável dependente (sem problemas computacionais), obtendo-se sempre um valor de GAIC para cada uma delas (Tabela 4.12). A distribuição *BE - beta* apresentou menor valor de GAIC, se ajustando melhor aos

dados para os dois tipos de infarto, portanto esta distribuição será utilizada para prosseguir com a modelagem.

Tabela 4.12: Comparação das distribuições ajustadas.

Infarto	Distribuiçao	Graus de liberdade	GAIC
Inferior	*Beta*	2	$-44,6803$
	Beta Generalizada Tipo 1	4	$-40,8235$
Anterior	*Beta*	2	$-53,2795$
	Beta Generalizada Tipo 1	4	$-50,0140$

ANÁLISE DO MODELO

A construção do modelo é efetuada com base nas variáveis de interesse, idade, sexo e D2 do infarto inferior, idade, sexo V1, V3, V4 e V6 do infarto anterior, para cada parâmetro da distribuição e em cada banco de dados. Tendo escolhido a distribuição, primeiramente foram ajustados modelos considerando o máximo de variável dependente para diferentes combinações em relação a utilização do suavizador *cubic spline*.

Nas Tabelas 4.13 e 4.14 temos os modelos testados, para o infarto inferior e anterior repectivamente, tanto para o parâmetro μ quanto para σ^2, assim como seus valores de GAIC e R^2 generalizado para comparação dos mesmos. As combinações onde não houve convergência foram desconsideradas na construção da tabela. Neste ponto seria possível já selecionar o modelo com menor valor de GAIC, mas decidiu-se por seguir com todos na modelagem.

Tabela 4.13: Modelos GAMLSS testados para o infarto inferior que não obtiveram problemas computacionais (modelando μ e σ^2).

Modelo	Forma funcional	GAIC	R^2 gen.	Considerações
1	$idade + sexo + D2$	$-41,5218$	$0,2178$	Modelo considerando todas as variáveis, sem a utilização de suavizadores.
2	$cs(idade) + sexo + cs(D2)$	$-67,3986$	$0,8043$	Modelo considerando todas as variáveis, com a utilização de suavizador para as variáveis idade e D2.
3	$idade + sexo + cs(D2)$	$-39,8128$	$0,4122$	Modelo considerando todas as variáveis, com a utilização de suavizador para a variável D2.
4	$cs(idade) + sexo + D2$	$-41,2166$	$0,4348$	Modelo considerando todas as variáveis, com a utilização de suavizador para as variável idade.

Tabela 4.14: Modelos GAMLSS testados para o infarto anterior que não obtiveram problemas computacionais (modelando μ e σ^2).

Modelo	Forma funcional	GAIC	R^2 gen.	Considerações
1	$idade+sexo+V1+V2+V3+V4+V6$	$-63,2127$	$0,7229$	Modelo considerando todas as variáveis, sem a utilização de suavizadores.
2	$idade+sexo+cs(V1)+V3+V4+V6$	$-65,3278$	$0,8202$	Modelo considerando todas as variáveis, com a utilização de suavizador para a variável V1.
3	$idade+sexo+V1+V3+cs(V4)+V6$	$-91,9299$	$0,9305$	Modelo considerando todas as variáveis, com a utilização de suavizador para a variável V4.

Tabela 4.15: Valores de GAIC para os modelos GAMLSS após utilização da função **stepGAICALL.A**.

Infarto	Modelo	GAIC	R^2 gen.
Inferior	1	$-42,7055$	$0,1058$
	2	$-69,8949$	$0,8069$
	3	$-44,5341$	$0,3194$
	4	$-45,1859$	$0,4343$
Anterior	1	$-64,6839$	$0,7176$
	2	$-86,8073$	$0,9103$
	3	$-100,9797$	$0,9419$

Na próxima etapa, com o auxílio da função **stepGAICALL.A**, para todos os modelos das Tabelas 4.13 e 4.14, foi construído um modelo para cada parâmetro da distribuição (Tabela 4.15), destes, para o infarto inferior, o modelo 2 apresentou menor valor de GAIC e maior de R^2 generalizado enquanto que, para o infarto anterior, obedecendo os mesmos critérios, o modelo selecionado foi o 3. A expressão dos modelos GAMLSS multivariável obtidos foram, para o infarto inferior:

$$g_1(\mu) = \beta_{10} + \beta_{11}cs(idade) + \beta_{13}cs(D2), \tag{4.3}$$

$$g_2(\sigma) = \beta_{20} + \beta_{21}cs(idade) + \beta_{22}sexo + \beta_{23}cs(D2). \tag{4.4}$$

para o infarto anterior:

$$g_1(\mu) = \beta_{10} + \beta_{12}sexo + \beta_{13}V1 + \beta_{14}V3 + \beta_{15}cs(V4), \tag{4.5}$$

$$g_2(\sigma) = \beta_{20} + \beta_{21}idade + \beta_{22}sexo + \beta_{23}V1 + \beta_{24}V3 + \beta_{25}cs(V4) + \beta_{26}V6, \tag{4.6}$$

Nas Tabelas 4.16 e 4.17 temos os valores das estimativas, erro padrão e valor p, onde podemos verificar que, para o modelo do infarto inferior, o valor p da variável idade para o parâmetro σ^2 é superior a $0,05$ (Tabela 4.16), indicando que não é rejeitada a possibilidade de o coeficiente ser igual a zero, portanto, pode ser considerada como pouco relevante para este modelo de predição. O mesmo acontece com as variáveis idade e V1 para o modelo do infarto anterior para o parâmetro σ^2 (Tabela 4.17).

Tabela 4.16: Estimativas dos parâmetros para $g(\mu)$ e $g(\sigma)$ para o modelo GAMLSS do infarto inferior.

Parâmetro modelado	**Variável**	**Estimativa**	**Erro Padrão**	**valor p**
μ	Intercepto	$-2,2373$	$0,0053$	$0,0000$
	cs(idade)	$0,0180$	$0,0001$	$0,0000$
	cs(D2)	$0,0648$	$0,0004$	$0,0000$
σ^2	Intercepto	$-11,2213$	$0,7663$	$0,0000$
	cs(idade)	$0,1799$	$0,0132$	$0,0000$
	sexo	$-0,6159$	$0,3131$	$0,0599$
	cs(D2)	$-0,5811$	$0,0939$	$0,0000$

Tabela 4.17: Estimativas dos parâmetros para $g(\mu)$ e $g(\sigma)$ para o modelo GAMLSS do infarto anterior.

Parâmetro modelado	**Variável**	**Estimativa**	**Erro Padrão**	**valor p**
μ	Intercepto	$-0,2045$	$0,0004$	$0,0000$
	sexo	$-0,1316$	$0,0004$	$0,0000$
	V1	$0,0203$	0.0004	$0,0000$
	V3	$-0,0711$	$0,0000$	$0,0000$
	cs(V4)	$0,0552$	0.0001	$0,0000$
σ^2	Intercepto	$-3,0306$	$1,0317$	0.0149
	idade	$-0,0268$	$0,0207$	$0,2231$
	sexo	$2,7253$	$0,6020$	$0,0011$
	V1	$-0,0845$	$0,2411$	$0,7334$
	V3	$-0,2733$	$0,0891$	$0,0119$
	cs(V4)	$1,0357$	$0,1630$	$0,0001$
	V6	$-2,0122$	$0,3169$	$0,0001$

ANÁLISE DE RESÍDUOS

Para verificar se os resíduos seguem distribuição normal foram construídos *box-plot* e *QQ-plot*. Na Figura 4.10 estão representados os gráficos para análise dos resíduos obtidos pelo modelo do infarto inferior. Como é possível observar, ambos os gráficos sugerem que os resíduos seguem distribuição normal. No gráfico *QQ-plot* (Figura 4.10) é possível observar que todos os resíduos estão próximos da reta diagonal ($y = x$), além disso, no gráfico *box-plot* nenhum ponto aparece como extremo.

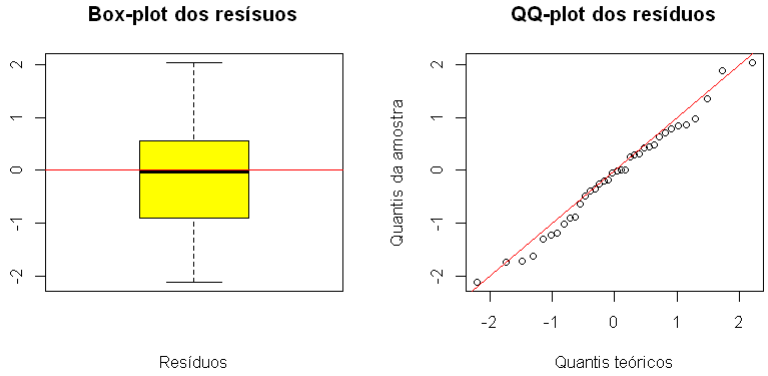

Figura 4.10: Box-plot (esquerda) e *QQ-plot* (direita) dos resíduos do modelo GAMLSS para o infarto inferior.

Na Figura 4.11 estão representados os gráficos para análise dos resíduos obtidos pelo modelo do infarto anterior. Como é possível observar, ambos os gráficos sugerem que os resíduos seguem distribuição normal. No gráfico *QQ-plot* (Figura 4.10) é possível observar que todos os resíduos estão próximos da reta diagonal ($y = x$), mas, no gráfico *box-plot*, um ponto aparece como valor extremo.

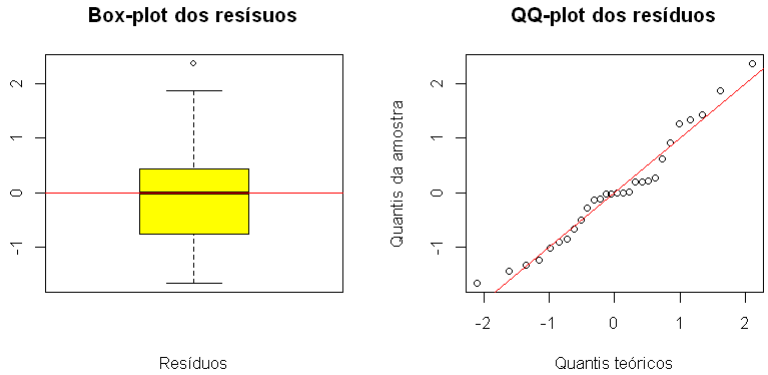

Figura 4.11: Box-plot (esquerda) e *QQ-plot* (direita) dos resíduos do modelo GAMLSS para o infarto anterior.

As medidas descritivas dos resíduos, obtidas através da função **plot()**, estão dispostas na Tabela 4.18, onde podemos ver que os resíduos têm uma média de $-0,1097$ e variância de $1,0020$ para o modelo do infarto inferior e uma média de $0,0196$ e variância de $1,0205$ para o modelo do infarto anterior, o que sugere que os resíduos para os dois modelos seguem distribuição normal padrão.

Tabela 4.18: Medidas descritivas dos resíduos do modelo GAMLSS ajustado.

Infarto	Medida	Valor
Inferior	Média	$-0,1097$
	Variância	$1,0020$
	coef. de assimetria	$0,0048$
	coef. de curtose	$2,4103$
	coef. de correlação de Filliben	$0,9938$
Anterior	Média	$0,0196$
	Variância	$1,0205$
	coef. de assimetria	$0,4197$
	coef. de curtose	$2,5324$
	coef. de correlação de Filliben	$0,9829$

Nas Figuras 4.12 e 4.13 estão apresentados os gráficos obtidos pela função **plot()**. Observando os gráficos do canto superior direito e esquerdo, não existe qualquer padrão para os resíduos, o que indica um bom ajustamento para o modelo. No gráfico localizado no canto inferior esquerdo das figuras, podemos observar que tem uma forma semelhante à da função densidade da normal padrão. Através dos testes F, Durbin-Watson e Shapiro-Wilk foram testadas as hipóteses de que os resíduos são homocedásticos, independentes e seguem distribuição normal, respectivamente, onde, ao nível de 5% de significância, nenhuma foi rejeitada, como pode ser visto na Tabela 4.19.

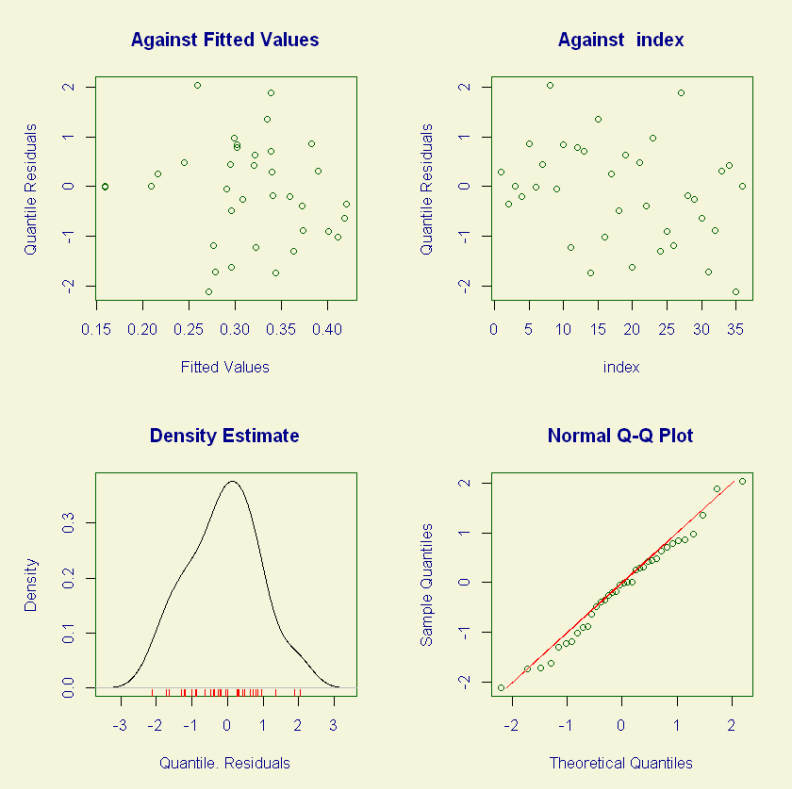

Figura 4.12: Gráficos dos resíduos obtidos através da função **plot()** do modelo GAMLSS para o infarto inferior.

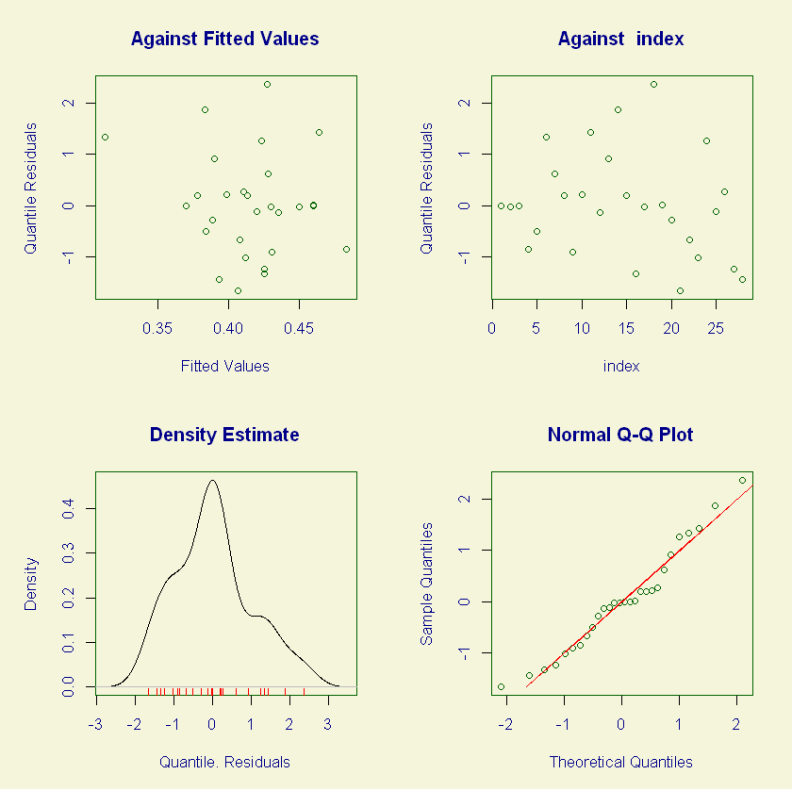

Figura 4.13: Gráficos dos resíduos obtidos através da função **plot()** do modelo GAMLSS para o infarto anterior.

Tabela 4.19: Testes para normalidade e homocedasticidade dos resíduos do modelo GAMLSS ajustado.

Infarto	**Teste**	**Hipótese nula**	**valor p**
Inferior	Shapiro-Wilk para normalidade	Os resíduos seguem distribuição normal	$0,8814$
	Teste F para homoscedasticidade	Os resíduos são homocedásticos	$0,5881$
	Durbin-Watson para independência	Os resíduos são independentes	$0,9911$
Anterior	Shapiro-Wilk para normalidade	Os resíduos seguem distribuição normal	$0,4095$
	Teste F para homoscedasticidade	Os resíduos são homocedásticos	$0,2846$
	Durbin-Watson para independência	Os resíduos são independentes	$0,1417$

Na Figura 4.14 estão representados os gráficos *worm plot* dos modelos GAMLSS para o infarto inferior (esquerda) e anterior (direita), onde pode-se verificar que todos pontos estão plotados dentro da banda de confiança de 95% e próximos da reta $y = 0$, indicando um bom ajuste do modelo.

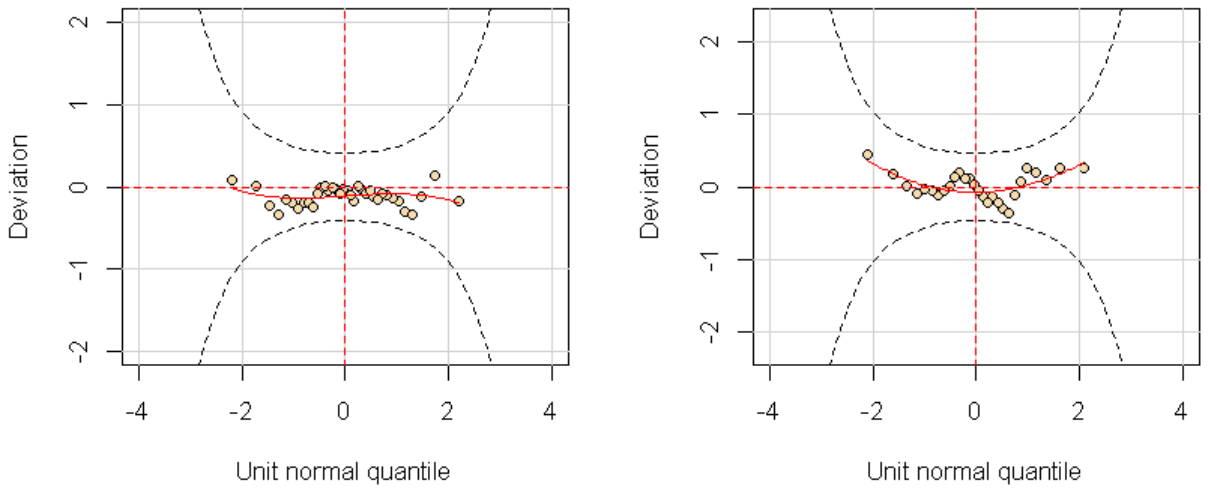

Figura 4.14: Gráficos *worm plot* dos modelos GAMLSS para o infarto inferior (esquerda) e anterior (direita).

5. Conclusões

Neste trabalho foram apresentadas características da família GAMLSS, com aplicação em um banco de dados analisado anteriormente, onde alguns registros foram excluídos. Na análise considerada usando GAMLSS nenhuma observação foi excluída. Através dos resultados obtidos pelo modelo GAMLSS foram encontrados modelos com valores de GAIC inferiores e R^2 generalizado superiores aos modelos de regressão beta para os dois tipos de infarto, sem a exclusão de observações do banco de dados, além de terem melhores resultados na análise de resíduos em relação às pressuposições de normalidade e homocedasticidade e também na avaliação do *worm plot*, alcançando o objetivo principal do trabalho.

Conclui-se então que os modelos GAMLSS são uma ferramenta poderosa no ajuste de modelos, uma vez que auxilia o pesquisador em etapas que muitas vezes são feitas de modo intuitivo, como por exemplo a escolha da melhor distribuição para determinado banco de dados, além de encontrar o melhor modelo para a situação proposta de modo rápido e com certa facilidade em relação a linguagem de programação. Outro ponto relavante é o fato de podermos modelar não apenas a média, como é feito nos GLM e GAM, mas também a dispersão, a assimetria e a curtose, abrindo mais possibilidades para o pesquisador e, consequentemente, atingindo melhores ajustes no processo de modelagem.

Referências Bibliográficas

[1] AKAIKE, H.: *Information measures and model selection.* Bulletin of the International Statistical Institute, 44:277–290, 1982.

[2] BUUREN, S. V. e FREDRIKS, M.: *Worm plot: a simple diagnostic device for modelling growth reference curves.* Statistics in medicine, 20(8):1259–1277, 2001. https://onlinelibrary.wiley.com/doi/10.1002/sim.746.

[3] COLE, T. J. e GREEN, P. J.: *Smoothing reference centile curves: the LMS method and penalized likelihood.* Statistics in medicine, 11(10):1305–1319, 1992. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780111005.

[4] DUNN, P. K. e SMYTH, G. K.: *Randomized quantile residuals.* Journal of Computational and Graphical Statistics, 5(3):236–244, 1996. https://www.jstor.org/stable/1390802?seq=1#page_scan_tab_contents.

[5] FERRARI, S. e CRIBARI-NETO, F.: *Beta regression for modelling rates and proportions.* Journal of Applied Statistics, 31(7):799–815, 2004. https://www.tandfonline.com/doi/abs/10.1080/0266476042000214501.

[6] FLORENCIO, L.: *Engenharia de avaliações com base em modelos GAMLSS.* Dissertação de Mestrado, 2010.

[7] HASTIE, T. e TIBSHIRANI, R.: *Generalized Additive Models: Some Applications.* Journal of the American Statistical Association, 82(398):371–386, 1987. https://www.jstor.org/stable/2289439?seq=1#page_scan_tab_contents.

[8] NAGELKERKE, N. J.: *A note on a general definition of the coefficient of determination.* Biometrika, 78(3):691–692, 1991. https://www.cesarzamudio.com/uploads/1/7/9/1/17916581/nagelkerke_n.j.d._1991_-_a_note_on_a_general_definition_of_the_coefficient_of_determination.pdf.

[9] PAULA, G.: *Modelos de Regressão com Apoio Computacional.* IME/USP, 2004.

[10] PINTO, E. R., PEREIRA, L. A., RESENDE, L. O. e DESTRO FILHO, J. B.: *Modelos estatísticos para estimação da área miocárdica sob risco de necrose.* Revista Brasileira Biometria, 29(3):295–415, 2011. http://jaguar.fcav.unesp.br/RME/fasciculos/v29/v29_n3/indice_v29_n3.php.

[11] RIGBY, B., STASINOPOULOS, M., HELLER, G. e VOUDOURIS, V.: *The distribution toolbox of GAMLSS*. The GAMLSS Team, 2014.

[12] RIGBY, R. A. e STASINOPOULOS, D. M.: *A Semi-parametric Additive Model for Variance Heterogeneity.* Statistics and Computing, 6(1):57–65, 1996. https://link.springer.com/article/10.1007/BF00161574.

[13] RIGBY, R. A. e STASINOPOULOS, D. M.: *Generalized additive models for location, scale and shape.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3):507–554, 2005. https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9876.2005.00510.x.

[14] STASINOPOULOS, M. D., RIGBY, R. A., HELLER, G. Z., VOUDOURIS, V. e DE BASTIANI, F.: *Flexible Regression and Smoothing The GAMLSS packages in R.* 2017.

[15] TEAM, R. C.: *R language definition.* Vienna, Austria: R foundation for statistical computing., 2000.

A. Apêndice

Nesta seção serão expostos os comandos elaborados no *software R* utilizados neste trabalho.

```
### Carregamento de pacotes ###
require(gamlss)
require(nortest)
require(lmtest)

### Leitura dos dados ###
#Infarto Inferior
inf_inf <- read.table("C:/Users/cassio_alcantara/Documents/ESTATISTICA/TCC/
dados/infarto_inferior.csv", header = T, sep = ";")
dados_inf <- data.frame(inf_inf)
attach(dados_inf)
#Infarto Anterior
inf_ant <- read.table("C:/Users/cassio_alcantara/Documents/ESTATISTICA/TCC/
dados/infarto_anterior.csv", header = T, sep = ";")
dados_ant <- data.frame(inf_ant)
attach(dados_ant)

### Estudo exploratório dos dados ###

#Medidas descritivas variáveis contínuas
summary(dados_inf)
summary(dados_ant)

#Histograma e Box-plot
op <- par(mfrow=c(2,2))
hist(dados_inf$ARN, xlab = "ARN", ylab = "Frequências", main = "Infarto
Inferior- Histograma da ARN", col = "blue")
boxplot(dados_inf$ARN, main = "Infarto Inferior- Box-plot da ARN",
xlab = "ARN", col = "blue")
hist(dados_ant$ARN, xlab = "ARN", ylab = "Frequências", main = "Infarto
Anterior- Histograma da ARN", col = "blue")
```

```
boxplot(dados_ant$ARN, main = "Infarto Anterior- Box-plot da ARN",
xlab = "ARN", col = "blue")
par(op)

#Infarto Inferior - Sexo
table(dados_inf$sexo)
prop.table(table(dados_inf$sexo))

#Infarto anterior - Sexo
table(dados_ant$sexo)
prop.table(table(dados_ant$sexo))

#Histograma e Box-plot
dadosm1 <- subset(dados_inf, sexo==1)
dadosh1 <- subset(dados_inf, sexo==0)
dadosm2 <- subset(dados_ant, sexo==1)
dadosh2 <- subset(dados_ant, sexo==0)
#infarto inferior
op <- par(mfrow=c(2,2))
boxplot(dadosh1$ARN, main = "Box-plot da ARN para os homens", xlab = "ARN"
, col = "blue")
boxplot(dadosm1$ARN, main = "Box-plot da ARN para as mulheres", xlab = "ARN"
, col = "blue")
par(op)
#infarto anterior
op <- par(mfrow=c(2,2))
boxplot(dadosh2$ARN, main = "Box-plot da ARN para os homens", xlab = "ARN"
, col = "blue")
boxplot(dadosm2$ARN, main = "Box-plot da ARN para as mulheres", xlab = "ARN"
, col = "blue")
par(op)

#Relação entre as variáveis
cor(dados_inf)
cor(dados_ant)

### Modelagem GAMLSS ###
#utilizando as funções fitdist e histDist para identificar possíveis funções
de ligação para a variável resposta:
```

```
fitdist1 <- fitDist(dados_inf$ARN, type = "real0to1")
fitdist1$fits
fitdist2 <- fitDist(dados_ant$ARN, type = "real0to1")
fitdist2$fits
GAIC(mBE1,mGB11,mBE2,mGB12)

op <- par(mfrow=c(2,2))
mBE1 <- histDist(ARN,family="BE", data=dados_inf, density = TRUE, main=
"Infarto Inferior - Distribuição BE") #beta
mGB11 <- histDist(ARN,family="GB1", data=dados_inf, density = TRUE, main=
"Infarto Inferior - Distribuição GB1") #beta
mBE2 <- histDist(ARN,family="BE", data=dados_ant, density = TRUE, main=
"Infarto Anterior - Distribuição BE") #beta
mGB12 <- histDist(ARN,family="GB1", data=dados_ant, density = TRUE, main=
"Infarto Anterior - Distribuição GB1") #beta
par(op)

## SELEÇÃO DE VARIÁVEIS PARA O MODELO - INFARTO INFERIOR

# modelo sem adição de cubic splines:
mbe0 <- gamlss(ARN~idade+sexo+D2, sigma.fo=~idade+sexo+D2, family = BE,
data = dados_inf,trace=FALSE)

# modelo com cubic splines para idade e D2:
mbe1 <- gamlss(ARN~cs(idade)+sexo+cs(D2), sigma.fo=~cs(idade)+sexo+cs(D2),
family = BE, data = dados_inf,trace=FALSE)

# modelo com cubic splines para D2:
mbe2 <- gamlss(ARN~idade+sexo+cs(D2), sigma.fo=~idade+sexo+cs(D2),
family = BE, data = dados_inf,trace=FALSE)

# modelo com cubic splines para idade:
mbe3 <- gamlss(ARN~cs(idade)+sexo+D2, sigma.fo=~cs(idade)+sexo+D2,
family = BE,data = dados_inf,trace=FALSE)

#GAIC dos modelos para comparação:
GAIC(mbe,mbe1,mbe2,mbe3)

#Valores do R-squared generalizado
Rsq(mbe0)
```

```
Rsq(mbe1)
Rsq(mbe2)
Rsq(mbe3)

# stepGAICALL.A
mbe4 <- stepGAICAll.A(mbe0, lower=~1, upper=~idade+sexo+D2,
steps = 1000000)
mbe5 <- stepGAICAll.A(mbe1, lower=~1, upper=~cs(idade)+sexo+cs(D2),
steps = 1000000)
mbe6 <- stepGAICAll.A(mbe2, lower=~1, upper=~idade+sexo+cs(D2),
steps = 1000000)
mbe7 <- stepGAICAll.A(mbe3, lower=~1, upper=~cs(idade)+sexo+D2,
steps = 1000000)

#GAIC dos modelos para comparação:
GAIC(mbe4,mbe5,mbe6,mbe7)

#Valores do R-squared generalizado
Rsq(mbe4)
Rsq(mbe5)
Rsq(mbe6)
Rsq(mbe7)

## SELEÇÃO DE VARIÁVEIS PARA O MODELO -INFARTO ANTERIOR

# modelo sem adição de cubic splines:
mbe8 <- gamlss(ARN~idade+sexo+V1+V2+V3+V4+V6, sigma.fo=~idade+sexo+V1+V3+V4+
V6,family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade, V1, V3 e V4:
mbe9 <- gamlss(ARN~cs(idade)+sexo+cs(V1)+cs(V3)+cs(V4)+V6, sigma.fo=~
cs(idade)+sexo+cs(V1)+cs(V3)+cs(V4)+V6, family = BE, data = dados_ant,
trace=FALSE)

# modelo com cubic splines para as variáveis idade, V1, V3 e V6:
mbe10 <- gamlss(ARN~cs(idade)+sexo+cs(V1)+cs(V3)+V4+cs(V6), sigma.fo=~
cs(idade)+sexo+cs(V1)+cs(V3)+V4+cs(V6), family = BE, data = dados_ant,
trace=FALSE)

# modelo com cubic splines para as variáveis idade, V1, V4 e V6:
```

```
mbe11 <- gamlss(ARN~cs(idade)+sexo+cs(V1)+V3+cs(V4)+cs(V6), sigma.fo=~
cs(idade)+sexo+cs(V1)+V3+cs(V4)+cs(V6), family = BE, data = dados_ant,
trace=FALSE)

# modelo com cubic splines para as variáveis idade, V3, V4 e V6:
mbe12 <- gamlss(ARN~cs(idade)+sexo+V1+cs(V3)+cs(V4)+cs(V6), sigma.fo=
~cs(idade)+sexo+V1+cs(V3)+cs(V4)+cs(V6), family = BE, data = dados_ant,
trace=FALSE)

# modelo com cubic splines para as variáveis V1, V3, V4 e V6:
mbe13 <- gamlss(ARN~idade+sexo+cs(V1)+cs(V3)+cs(V4)+cs(V6), sigma.fo=~idade+
sexo+cs(V1)+cs(V3)+cs(V4)+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade, V1 e V3:
mbe14 <- gamlss(ARN~cs(idade)+sexo+cs(V1)+cs(V3)+V4+V6, sigma.fo=~cs(idade)+
sexo+cs(V1)+cs(V3)+V4+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade, V1 e V4:
mbe15 <- gamlss(ARN~cs(idade)+sexo+cs(V1)+V3+cs(V4)+V6, sigma.fo=~cs(idade)+
sexo+cs(V1)+V3+cs(V4)+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade, V1 e V6:
mbe16 <- gamlss(ARN~cs(idade)+sexo+cs(V1)+V3+V4+cs(V6), sigma.fo=~cs(idade)+
sexo+cs(V1)+V3+V4+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade, V3 e V4:
mbe17 <- gamlss(ARN~cs(idade)+sexo+V1+cs(V3)+cs(V4)+V6, sigma.fo=~cs(idade)+
sexo+V1+cs(V3)+cs(V4)+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade, V3 e V6:
mbe18 <- gamlss(ARN~cs(idade)+sexo+V1+cs(V3)+V4+cs(V6), sigma.fo=~cs(idade)+
sexo+V1+cs(V3)+V4+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade, V4 e V6:
mbe19 <- gamlss(ARN~cs(idade)+sexo+V1+V3+cs(V4)+cs(V6), sigma.fo=~cs(idade)+
sexo+V1+V3+cs(V4)+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis V1, V3 e V4:
mbe20 <- gamlss(ARN~idade+sexo+cs(V1)+cs(V3)+cs(V4)+V6, sigma.fo=~idade+sexo+
cs(V1)+cs(V3)+cs(V4)+V6, family = BE, data = dados_ant,trace=FALSE)
```

```
# modelo com cubic splines para as variáveis V1, V3 e V6:
mbe21 <- gamlss(ARN~idade+sexo+cs(V1)+cs(V3)+V4+cs(V6), sigma.fo=~idade+sexo+
cs(V1)+cs(V3)+V4+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis V3, V4 e V6:
mbe22 <- gamlss(ARN~idade+sexo+V1+cs(V3)+cs(V4)+cs(V6), sigma.fo=~idade+sexo+
V1+cs(V3)+cs(V4)+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade e V1:
mbe23 <- gamlss(ARN~cs(idade)+sexo+cs(V1)+V3+V4+V6, sigma.fo=~cs(idade)+sexo+
cs(V1)+V3+V4+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade e V3:
mbe24 <- gamlss(ARN~cs(idade)+sexo+V1+cs(V3)+V4+V6, sigma.fo=~cs(idade)+sexo+
V1+cs(V3)+V4+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade e V4:
mbe25 <- gamlss(ARN~cs(idade)+sexo+V1+V3+cs(V4)+V6, sigma.fo=~cs(idade)+sexo+
V1+V3+cs(V4)+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis idade e V6:
mbe26 <- gamlss(ARN~cs(idade)+sexo+V1+V3+V4+cs(V6), sigma.fo=~cs(idade)+sexo+
V1+V3+V4+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis V1 e V3:
mbe27 <- gamlss(ARN~idade+sexo+cs(V1)+cs(V3)+V4+V6, sigma.fo=~idade+sexo+
cs(V1)+cs(V3)+V4+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis V1 e V4:
mbe28 <- gamlss(ARN~idade+sexo+cs(V1)+V3+cs(V4)+V6, sigma.fo=~idade+sexo+
cs(V1)+V3+cs(V4)+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis V1 e V6:
mbe29 <- gamlss(ARN~idade+sexo+cs(V1)+V3+V4+cs(V6), sigma.fo=~idade+sexo+
cs(V1)+V3+V4+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis V3 e V4:
mbe30 <- gamlss(ARN~idade+sexo+V1+cs(V3)+cs(V4)+V6, sigma.fo=~idade+sexo+V1+
cs(V3)+cs(V4)+V6, family = BE, data = dados_ant,trace=FALSE)
```

```
# modelo com cubic splines para as variáveis V3 e V6:
mbe31 <- gamlss(ARN~idade+sexo+V1+cs(V3)+V4+cs(V6), sigma.fo=~idade+sexo+V1+
cs(V3)+V4+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para as variáveis V4 e V6:
mbe32 <- gamlss(ARN~idade+sexo+V1+V3+cs(V4)+cs(V6), sigma.fo=~idade+sexo+V1+
V3+cs(V4)+cs(V6), family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para a variável idade:
mbe33 <- gamlss(ARN~cs(idade)+sexo+V1+V3+V4+V6, sigma.fo=~cs(idade)+sexo+V1+
V3+V4+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para a variável V1:
mbe34 <- gamlss(ARN~idade+sexo+cs(V1)+V3+V4+V6, sigma.fo=~idade+sexo+cs(V1)+
V3+V4+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para a variável V3:
mbe35 <- gamlss(ARN~idade+sexo+V1+cs(V3)+V4+V6, sigma.fo=~idade+sexo+V1+
cs(V3)+V4+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para a variável V4:
mbe36 <- gamlss(ARN~idade+sexo+V1+V3+cs(V4)+V6, sigma.fo=~idade+sexo+V1+V3+
cs(V4)+V6, family = BE, data = dados_ant,trace=FALSE)

# modelo com cubic splines para a variável V6:
mbe37 <- gamlss(ARN~idade+sexo+V1+V3+V4+cs(V6), sigma.fo=~idade+sexo+V1+V3+V4+
cs(V6), family = BE, data = dados_ant,trace=FALSE)

#GAIC dos modelos para comparação:
GAIC(mbe8,mbe34,mbe36)

#Valores do R-squared generalizado
Rsq(mbe8)
Rsq(mbe34)
Rsq(mbe36)

# stepGAICALL.A
mbe38 <- stepGAICAll.A(mbe8, lower=~1, upper=~idade+sexo+V1+V2+V3+V4+V6,
steps = 1000000)
```

```
mbe39 <- stepGAICAll.A(mbe34, lower=~1, upper=~idade+sexo+cs(V1)+V3+V4+V6,
steps = 1000000)
mbe40 <- stepGAICAll.A(mbe36, lower=~1, upper=~idade+sexo+V1+V3+cs(V4)+V6,
steps = 1000000)

#GAIC dos modelos para comparação:
GAIC(mbe38,mbe39,mbe40)

#Valores do R-squared generalizado
Rsq(mbe38)
Rsq(mbe39)
Rsq(mbe40)

# Modelos selecionados:
summary(mbe5)
summary(mbe40)

## ANÁLISE DOS RESÍDUOS
residuos1 <- residuals(mbe5)
residuos2 <- residuals(mbe40)

op <- par(mfrow=c(2,2))
boxplot(residuos1, col = 'yellow', main = 'Box-plot dos resísuos', xlab =
'Resíduos')
abline(h=0, col = 'red')
qqnorm(residuos1, main = 'QQ-plot dos resíduos', xlab = 'Quantis teóricos',
ylab = 'Quantis da amostra')
abline(0, 1, col = 'red')
par(op)

op <- par(mfrow=c(2,2))
boxplot(residuos2, col = 'yellow', main = 'Box-plot dos resísuos', xlab =
'Resíduos')
abline(h=0, col = 'red')
qqnorm(residuos2, main = 'QQ-plot dos resíduos', xlab = 'Quantis teóricos',
ylab = 'Quantis da amostra')
abline(0, 1, col = 'red')
par(op)

plot(mbe5)
```

```
plot(mbe40)

#teste de normalidade dos resíduos
shapiro.test(residuos1)
shapiro.test(residuos2)

#teste de homoscedasticidade dos resíduos
var.test(residuos1[residuos1>0],residuos1[residuos1<0])
var.test(residuos2[residuos2>0],residuos2[residuos2<0])

#teste de independência dos resíduos
dwtest(mbe5)
dwtest(mbe40)

#worm plot
wp(mbe5)
wp(mbe40)

##MODELO DE REGRESSÃO BETA SEM EXCLUSÃO DE OBSERVAÇÕES DO BANCO DE DADOS
betainf <- gamlss(ARN~idade+sexo+D2, family = BE, data = dados_inf,
trace=FALSE)
betaant <- gamlss(ARN~idade+sexo+V1+V3+V4+V6, family = BE, data =
dados_ant,trace=FALSE)

betainf2 <- stepGAICAll.A(betainf, lower=~1, upper=~idade+sexo+D2,
steps = 1000000)
betaant2 <- stepGAICAll.A(betaant, lower=~1, upper=~idade+sexo+V1+V3+V4+V6,
steps = 1000000)

#Valores do R-squared generalizado
Rsq(betainf)
Rsq(betainf2)
Rsq(betaant)
Rsq(betaant2)

#valores de GAIC
GAIC(betainf,betainf2)
GAIC(betaant,betaant2)

#summary dos melhores modelos
```

```
summary(betainf2)
summary(betaant2)

## ANÁLISE DOS RESÍDUOS
residuos_inf <- residuals(betainf2)
residuos_ant <- residuals(betaant2)

op <- par(mfrow=c(2,2))
boxplot(residuos_inf, col = 'yellow', main = 'Box-plot dos resísuos', xlab =
'Resíduos')
abline(h=0, col = 'red')
qqnorm(residuos_inf, main = 'QQ-plot dos resíduos', xlab = 'Quantis teóricos',
ylab = 'Quantis da amostra')
abline(0, 1, col = 'red')
par(op)

op <- par(mfrow=c(2,2))
boxplot(residuos_ant, col = 'yellow', main = 'Box-plot dos resísuos', xlab =
'Resíduos')
abline(h=0, col = 'red')
qqnorm(residuos_ant, main = 'QQ-plot dos resíduos', xlab = 'Quantis teóricos',
ylab = 'Quantis da amostra')
abline(0, 1, col = 'red')
par(op)

plot(betainf2)
plot(betaant2)

#teste de normalidade dos resíduos
shapiro.test(residuos_inf)
shapiro.test(residuos_ant)

#teste de homoscedasticidade dos resíduos
var.test(residuos_inf[residuos_inf>0],residuos_inf[residuos_inf<0])
var.test(residuos_ant[residuos_ant>0],residuos_ant[residuos_ant<0])

#teste de independência dos resíduos
dwtest(betainf2)
dwtest(betaant2)
```

```
#worm plot
op <- par(mfrow=c(2,2))
wp(betainf2)
wp(betaant2)
par(op)
```

B. Apêndice

Tabela B.1: Dados sobre ARN para infarto inferior.

Obs	ARN	idade	sexo	D2	D3	aVF	Obs	ARN	idade	sexo	D2	D3	aVF
1	0.38	76	0	2.5	2	2.5	19	0.53	53	0	1	1	1
2	0.12	79	0	1	2	1.5	20	0.11	55	1	2	3	3
3	0.21	40	0	1.5	1.5	1	21	0.27	52	1	2.5	3.5	3
4	0.20	58	0	1	0.5	1.5	22	0.18	60	0	1	1.5	2
5	0.51	64	1	1.5	2	2	23	0.36	64	0	3	3	3
6	0.16	41	0	3	3	3	24	0.22	68	1	2	3	2.5
7	0.34	57	0	4.5	5	1	25	0.21	70	1	1.5	1	1
8	0.47	52	0	4.5	5	4.5	26	0.19	59	0	3.5	5	4
9	0.29	69	1	6.5	9	7.5	27	0.54	74	1	2.5	2	2.5
10	0.36	60	0	5.5	7	6.5	28	0.32	76	1	4	5	4
11	0.22	79	1	3	4.5	4	29	0.28	62	0	2.5	3	3
12	0.35	64	0	3.5	4	4	30	0.16	74	1	1	1.5	1.5
13	0.40	65	0	5	6.5	6	31	0.11	49	0	1.5	2.5	2
14	0.16	74	0	5.5	5	6	32	0.16	62	0	1.5	1.5	1
15	0.53	71	0	2.5	4	4.5	33	0.47	63	0	1	2	1
16	0.08	68	1	0.5	2.5	1.5	34	0.35	75	1	3	5	4
17	0.22	48	1	2.5	4	3	35	0.15	47	0	1	2	1.5
18	0.16	50	0	1	2	1	36	0.16	39	0	2.5	3	2.5

Tabela B.2: Dados sobre ARN para infarto anterior.

Obs	ARN	idade	sexo	V1	V2	V3	V4	V5	V6
1	0.46	73	0	2.5	5.5	8.5	8	4.5	1.5
2	0.43	54	1	1.5	4	1.5	0.5	0	0
3	0.37	65	1	2	5	5.5	0	0	0
4	0.39	89	0	2	5	4	6.5	3.5	0
5	0.24	86	0	2.5	4	7.5	5	2	0
6	0.41	55	0	1	7.5	13	6	3	0
7	0.48	73	0	2	2.5	2.5	2	0	0
8	0.45	68	0	1	2	3	2.5	1	0
9	0.35	68	0	2	5.5	7	6.5	4	0
10	0.42	57	0	1	3	4	3	2	1
11	0.54	76	0	2	3	3.5	5.5	4	1.5
12	0.40	48	0	0	2	1.5	2	0	0
13	0.60	65	1	0.5	1	2.5	3	2.5	2
14	0.53	87	1	2	4.5	3.5	1.5	0	0
15	0.41	62	0	3	6	8	5	1	0
16	0.38	61	0	0	4	2	2.5	1	1.5
17	0.45	53	0	2.5	5	2	1	1.5	1
18	0.48	50	0	1	2	2.5	1.5	0	0
19	0.46	67	0	1	2	1.5	0.5	0	0
20	0.21	48	0	2.5	7	5	3	1.5	0
21	0.29	61	0	3	4	4	2	0	0
22	0.32	46	0	1.5	3.5	3.5	2	0	0
23	0.33	48	1	0	0	1	2	2.5	2
24	0.45	57	0	1	2	2.5	2	1	1
25	0.42	66	0	0	3.5	3	1	0	0
26	0.50	67	0	0	2	3	3	0	0
27	0.40	57	0	1.5	3	2.5	2	1.5	1
28	0.32	71	0	1.5	8	9	6.5	2	0