UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
CURSO DE ESTATÍSTICA

Lucas Avila Moreira de Paula

Modelagem conjunta dos parâmetros de locação, escala, assimetria e curtose dos modelos skew-normal e skew-t utilizando o GAMLSS: aplicações na Economia, Medicina e Zoologia

Juiz de Fora
2024

Lucas Avila Moreira de Paula

Modelagem conjunta dos parâmetros de locação, escala, assimetria e curtose dos modelos skew-normal e skew-t utilizando o GAMLSS: aplicações na Economia, Medicina e Zoologia

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientadora: Prof[a]. Dra. Camila Borelli Zeller

Coorientador: Prof. Dr. João Henrique Gonçalves Mazzeu

Juiz de Fora

2024

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Paula, Lucas Avila Moreira de Paula.
Modelagem conjunta dos parâmetros de locação, escala, assimetria e curtose dos modelos skew-normal e skew-t utilizando o GAMLSS : aplicações na Economia, Medicina e Zoologia / Lucas Avila Moreira de Paula. – 2024.
68 f. : il.

Orientadora: Camila Borelli Zeller
Coorientador: João Henrique Gonçalves Mazzeu
Conclusão de Curso (graduação) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Curso de Estatística, 2024.

1. GAMLSS. 2. modelagem conjunta. 3. distribuições assimétricas. 4. distribuições simétricas. 5. dados reais. I. Zeller, Camila Borelli, orient. II. Mazzeu, Jõao Henrique Gonçalves, coorient. III. Título.

Lucas Avila Moreira de Paula

Modelagem conjunta dos parâmetros de locação, escala, assimetria e curtose dos modelos skew-normal e skew-t utilizando o GAMLSS: aplicações na Economia, Medicina e Zoologia

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Aprovada em (dia) de (mês) de (ano)

BANCA EXAMINADORA

Profª. Dra. Camila Borelli Zeller - Orientador
Universidade Federal de Juiz de Fora

Professor Dr. João Henrique Gonçalves Mazzeu - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Clécio da Silva Ferreira
Universidade Federal de Juiz de Fora

Prof. Dr. Tiago Maia Magalhães
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Agradeço às minhas avós Maria e Nair, aos meus pais, Marta e Sérgio, aos meus irmãos, Matheus, Thiago e Vinicius, às minhas tias Cea, Cecília, Cristina e Sandra, e a todos os meus primos e primas, que são muitos para listar, mas sem os quais nada disso seria possível.

Agradeço também aos meus amigos, em especial ao Chus, Davi, Enzo, Marcos, Matheus (o ruim), Rafael e Rique, por todo o apoio desde a escola; ao Bruno e ao ET, por terem sido meus guias; e também ao Bruno (dessa vez, o gremista), Ral e Álvaro, que mesmo fisicamente distantes, me ajudaram muito.

Agradeço ainda ao Cláudio, meu orientador na Embrapa, por todos os ensinamentos e ajudas, e à Camila e ao João, por toda paciência na construção deste trabalho.

RESUMO

O processo de modelagem de dados reais via modelo de regressão linear (MRL) pode-se tornar uma tarefa árdua quando a distribuição da variável resposta não é gaussiana (normal), podendo apresentar variabilidade não linear ou não constante. Devido a isso, na literatura estatística, foram propostos modelos alternativos, dentre os quais podemos citar os modelos lineares generalizados (MLG), os modelos aditivos generalizados (GAM) e os modelos aditivos generalizados para locação, escala e forma (GAMLSS). A vantagem do MLG e do GAM em relação ao MRL é que a variável resposta pode seguir qualquer distribuição pertencente à família exponencial. No GAMLSS, a suposição de que a variável resposta pertence à família exponencial é relaxada e substituída por uma família de distribuições mais geral incluindo distribuições contínuas com assimetria ou curtose acentuadas e distribuições discretas. Adicionalmente, a parte sistemática do GAMLSS permite que não apenas a medida de posição, mas todos os parâmetros da distribuição da variável resposta sejam modelados através de funções paramétricas ou não-paramétricas (de suavização) das variáveis explicativas e\ou termos de efeitos aleatórios. Neste trabalho, com o intuito de avaliar a flexibilidade da distribuição da variável resposta no GAMLSS paramétrico, serão utilizadas as distribuições simétricas, normal e t-Student, e as suas versões assimétricas, skew-normal e skew-t. A eficácia do GAMLSS será demonstrada através da análise de conjuntos de dados simulados e reais no contexto da Medicina, Economia e Zootecnia. A metodologia abordada neste trabalho está implementada no pacote gamlss do software R.

Palavras-chave: GAMLSS; modelagem conjunta; distribuições assimétricas; distribuições simétricas; dados reais.

ABSTRACT

Modeling real data via Linear Regression (LR) may be a difficult task when the response variable does not follow a normal distribution or may present non-linear or non-constant variability. Due to these problems, alternative models were proposed, among them we can cite the Generalized Linear Models (GLM), Generalized Additive Models (GAM) and the Generalized Additive Models for Location, Scale and Shape (GAMLSS). The GLM and GAM have the advantage over LR for requiring that the response variable follows any distribution of the exponential family. In GAMLSS, that requirement is no longer necessary, allowing continuous distributions that present asymmetry or even strong skewness, as well as discrete distributions. Additionally, the systematic part of GAMLSS allows not only the scale parameters, but also all the parameters of the response variable to be modeled through parametric or non-parametric (smoothing) functions and\or random effects terms. In this work, in order to assess the flexibility of the response variable in parametric GAMLSS, we will use the normal and t-Student distributions, as well as their asymmetrical versions, skew-normal and skew-t. The efficiency of GAMLSS will be demonstrated through real data analysis in different contexts, such as medicine, economy and zootechnics. The methodology applied in this work is implemented in the gamlss package of the R software.

Keywords: GAMLSS; joint modeling; asymmetrical distributions; symmetrical distributions; real data.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os modelos de regressão linear (MRL) propostos por Galton (1886) são muito úteis para modelagem de uma variável resposta que segue distribuição normal, com média e variância constantes, sendo explicada por covariáveis com as quais possui relação linear. Por muitos anos, 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.

Para ampliar as possibilidades de modelagem, Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MLG), nos quais relaxaram a suposição de normalidade para a variável resposta. Neste caso, a distribuição da variável resposta deve pertencer à família exponencial. Além disso, uma função de ligação é considerada para modelar a relação entre a variável resposta e as explicativas. Com o surgimento de técnicas de suavização, Hastie e Tibshirani (1990) propuseram os modelos aditivos generalizados (GAM), incorporando, assim, relações não lineares entre as variáveis explicativas e a resposta, mantendo, no entanto, a restrição para modelar apenas variáveis que sigam distribuições dessa mesma família.

Nos modelos aditivos generalizados para localização, escala e forma (GAMLSS: Generalized Additive Models for Location, Scale and Shape) apresentados por Rigby e Stasinopoulos (2005), a suposição de que a variável resposta pertence à família exponencial é relaxada e substituída por uma classe de distribuições mais geral incluindo distribuições contínuas com assimetria ou curtose acentuadas e distribuições discretas. Adicionalmente, o GAMLSS permite a modelagem conjunta da média, variância, assimetria e curtose em funções de covariáveis. Apesar do foco deste trabalho ser em modelos paramétricos, é importante salientar que o GAMLSS também permite a modelagem de termos não-paramétricos com o uso de funções de suavização nas covariáveis e termos aleatórios, o que mostra ainda mais a sua flexibilidade em modelar a distribuição da variável resposta. Neste trabalho, com o intuito de avaliar a flexibilidade da distribuição da variável resposta no GAMLSS paramétrico, serão utilizadas as distribuições simétricas, normal e t-Student, e as suas versões assimétricas, skew-normal e skew-t. A eficácia do GAMLSS será demonstrada através da análise de conjunto de dados simulados e reais no contexto da Medicina, Economia e Zootecnia.

O GAMLSS tem sido usado em uma variedade de campos aplicados, incluindo ciências sociais, ciências ambientais, finanças, ciências atuariais, biologia, biociências, energia, genômica, pesca, consumo de alimentos, estimativa de curva de crescimento, pesquisa marinha, Medicina, Meteorologia, chuvas, vacinas, etc. A título de exemplos específicos em algumas instituições internacionais importantes, Rigby et al. (2019) menciona

- A OMS usa o GAMLSS para construir gráficos de curvas de crescimento para crianças.

Esses gráficos são usados em mais de 140 países (World Health Organization, 2006);

- O FMI usa o GAMLSS para analisar fatores que possam causar risco sistêmico no sistema financeiro (International Monetary Fund, 2015).

Além disso, em Serinaldi (2011), o GAMLSS foi utilizado para previsão de preços da eletricidade, enquanto Ramires et al. (2019) utilizou para análise de sobrevivência no contexto de reincidência de crimes no Brasil e no contexto de câncer de mama na Alemanha.

1.1 OBJETIVOS

Neste trabalho, temos como objetivo estudar o GAMLSS, usando o software R (R Core Team, 2023) e, mais especificamente, o pacote gamlss implementado por Rigby e Stasinopoulos (2005). Nesse contexto, buscamos entender a estimação das relações paramétricas entre a variável resposta e as explicativas, a escolha das distribuições, a seleção das variáveis explicativas e o diagnóstico do modelo. Exemplos numéricos considerando dados simulados e reais serão apresentados. Na aplicação analisaremos, com o ferramental previamente descrito, dados reais no contexto da economia, onde serão analisadas relações entre taxas de retorno; da saúde, em que será avaliada a influência da idade gestacional na circunferência abdominal de fetos; e da zootecnia, em que buscaremos compreender como alguns fatores afetam a produção total de leite de vacas da raça Holandesa. Ao final deste estudo, espera-se que os leitores estejam familiarizados com os aspectos fundamentais do GAMLSS que acreditamos serem de grande aplicabilidade.

1.2 ESTRUTURA DO TRABALHO

Este trabalho encontra-se dividido em 5 capítulos. No Capítulo 2, temos a metodologia, abrangendo aspectos teóricos, de forma introdutória, do GAMLSS paramétrico que será aplicado no trabalho, no contexto de distribuições simétricas e assimétricas, incluindo discussões sobre a escolha da distribuição, seleção das variáveis explicativas, inferência, interpretação e diagnóstico do modelo. No Capítulo 3, exemplos numéricos considerando dados simulados serão apresentados para ilustrar o modelo e os resultados inferenciais desenvolvidos, comparando, quando possível, com os resultados alcançados por outros trabalhos. Seguindo, no Capítulo 4 será feita a aplicação da metodologia descrita no Capítulo 2, utilizando base de dados reais das áreas de Economia, da Medicina e da Zootecnia. Por fim, no Capítulo 5, concluiremos sobre os resultados obtidos no presente trabalho, além de sugerir trabalhos futuros.

2 ASPECTOS TEÓRICOS

Este capítulo apresenta a metodologia utilizada neste trabalho, detalhando os procedimentos e técnicas consideradas para se atingir os objetivos declarados na Seção 1.1. Inicialmente, abordaremos o GAMLSS que será aplicado e, posteriormente, serão apresentadas as técnicas de inferência e diagnóstico do modelo. Por fim, uma apresentação dos equipamentos e ferramentas utilizadas nos exemplos numéricos considerando dados simulados e reais descritos nos Capítulos 3 e 4.

2.1 GAMLSS PARAMÉTRICO

Em Rigby e Stasinopoulos (2005) é apresentada uma classe geral de modelos estatísticos para uma variável de resposta univariada que chamamos de modelo aditivo generalizado para localização, escala e forma (GAMLSS). O modelo pressupõe observações independentes da variável resposta Y dado os parâmetros, as variáveis explicativas e os valores dos efeitos aleatórios. A distribuição para a variável resposta no GAMLSS pode ser selecionada a partir de uma classe geral de distribuições, incluindo distribuições contínuas e discretas altamente assimétricas ou curtóticas. A parte sistemática do modelo é expandida para permitir a modelagem não apenas da média (ou localização), mas também dos demais parâmetros da distribuição de Y. Segue abaixo a definição do GAMLSS paramétrico, foco deste trabalho.

Considere $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ o vetor de observações realizadas de uma variável resposta com função densidade de probabilidade dada por $f(y|\boldsymbol{\theta})$ e o vetor paramétrico $\boldsymbol{\theta}^T = (\theta_1, \theta_2, \ldots, \theta_p)$, em que as observações y_i para $i = 1, 2, \ldots, n$ são independentes e condicionais a $\boldsymbol{\theta^i}$, com função densidade de probabilidade (ou função de massa de probabilidade no caso discreto) $f(y_i|\boldsymbol{\theta^i})$ e o vetor paramétrico $\boldsymbol{\theta^i} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})$ relacionado às variáveis explicativas e efeitos aleatórios.

Ao se considerar uma distribuição de probabilidade com quatro parâmetros para modelar uma variável resposta Y, um GAMLSS paramétrico pode ser escrito como

$$Y \sim \mathcal{D}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\theta_3}, \boldsymbol{\theta_4}), \tag{2.1}$$

e cada parâmetro distribucional estimado por um preditor específico, tal que

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k, \tag{2.2}$$

em que $k = 1, \ldots, \{1, 2, 3, 4\}$, $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores de tamanho n, isto é, $\boldsymbol{\theta_k^T} = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})$, $\boldsymbol{\beta_k^T} = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{d_k k})$ é o vetor de parâmetro de efeitos fixos de dimensão d_k, $\boldsymbol{X_k^T}$ é a matriz $n \times d_k$ de delineamento incorporando, de modo aditivo, termos lineares à parte

Parameter range	Link functions	Formula for $g(\theta)$
$-\infty$ to ∞	`identity`	θ
0 to ∞	`log`	$\log(\theta)$
	`sqrt`	$\sqrt{\theta}$
	`inverse`	$1/\theta$
	`'1/mu^2'`	$1/\theta^2$
	`'mu^2'`	θ^2
0 to 1	`logit`	$\log[\theta/(1-\theta)]$
	`probit`	$\Phi^{-1}(\theta)$
	`cauchit`	$\tan(\pi(\theta-0.05))$
	`cloglog`	$\log(-\log(1-\theta))$
1 to ∞	`logshiftto1`	$\log(\theta-1)$
2 to ∞	`logshiftto2`	$\log(\theta-2)$
0.00001 to ∞	`logshiftto0` or `Slog`[1]	$\log(\theta-0.00001)$

Figura 1: Funções de ligação implementadas no pagote gamlss.

paramétrica do modelo, e $g_k(.)$ é a função de ligação dos parâmetros de distribuição ao preditor linear $\boldsymbol{\eta_k}$. No pacote gamlss do R, há várias funções de ligação (as possibilidades já implementadas estão listadas na Figura 1 retirada de Rigby et al. (2019), pág. 174), mas usaremos apenas a identidade e log. Os dois primeiros parâmetros populacionais $\boldsymbol{\theta_1}$ e $\boldsymbol{\theta_2}$ no modelo definido em (2.1) e (2.2) são geralmente caracterizados como parâmetros de locação e escala, denotados aqui por $\boldsymbol{\mu}$ e $\boldsymbol{\sigma}$, enquanto que os parâmetros restantes são caracterizados como parâmetros de forma, por exemplo, $\boldsymbol{\theta_3} = \boldsymbol{\lambda}$ e $\boldsymbol{\theta_4} = \boldsymbol{\tau}$ são parâmetros de assimetria e curtose, respectivamente.

Por exemplo, Y $\sim \mathcal{D}(logit(\mu) = x, log(\sigma) = x, log(\lambda) = 1)$ é um GAMLSS em que a variável resposta Y tem distribuição $\mathcal{D}$ (conhecida), o parâmetro de posição é modelado usando uma função logito em x, o parâmetro σ é modelado a partir de um modelo log-linear em x e o parâmetro λ admitimos como constante e igual a 1, mas na escala logarítmica.

2.2 DISTRIBUIÇÕES DE INTERESSE

Para a modelagem dos dados usando o GAMLSS, qualquer distribuição paramétrica pode ser utilizada. Isto significa maior liberdade na criação de modelos de regressão, mas também mais responsabilidade na decisão de qual distribuição utilizar. Veja mais detalhes no pacote gamlss do software R. No software R, a única restrição para a implementação dos modelos GAMLSS é que as primeiras derivadas de $f(y|\boldsymbol{\theta})$, com relação aos parâmetros $\boldsymbol{\theta}$, sejam calculáveis. Derivadas explícitas são preferíveis, mas é possível utilizar funções numéricas para o cálculo dessas derivadas. Contudo, focaremos em quatro distribuições específicas, as distribuições simétricas, normal e t-Student, e as suas versões assimétricas, skew-normal e skew-t. Note que estas distribuições são frequentemente utilizadas na literatura estatística para modelagem de dados simétricos/assimétricos e/ou caudas pesadas. Neste trabalho, revisitamos estas distribuições no contexto do GAMLSS, motivados pelos trabalhos de Taylor e Verbyla (2004) e Li e Wu (2014), por exemplo.

2.2.1 Skew-normal

Em muitas análises na literatura estatística, existe uma tendência geral de se supor a normalidade dos dados, algo que nem sempre é o mais adequado. Por conta disso, nos últimos anos o estudo de modelos mais flexíveis ao gaussiano tem sido de grande interesse dos pesquisadores. Sob essa motivação é que será proposto, neste trabalho, um estudo baseado na distribuição skew-normal. Esta distribuição inclui a distribuição normal, como caso especial ($\lambda = 0$), e fornece flexibilidade em capturar uma ampla variedade de comportamentos não normais, por simplesmente adicionar um parâmetro que controla o grau de assimetria.

A distribuição skew-normal utilizada é a proposta por Azzalini (1985) e tem a seguinte função densidade de probabilidade

$$f(y|\mu,\sigma^2,\lambda) = \frac{2}{\sigma}\phi\left(\frac{(y-\mu)}{\sigma}\right)\Phi\left(\lambda\frac{(y-\mu)}{\sigma}\right), y \in \mathbb{R}, \tag{2.3}$$

onde $\phi(.)$ e $\Phi(.)$ são, respectivamente, a função densidade de probabilidade e a função de distribuição acumulada da normal padrão, $-\infty < \mu < \infty$, $\sigma > 0$ e $-\infty < \lambda < \infty$. Importante ressaltar que λ caracteriza a assimetria da densidade, sendo $\lambda > 0$ assimetria positiva, $\lambda < 0$ assimetria negativa e $\lambda = 0$ simetria, o que retorna a densidade da normal. Para essa distribuição, temos que

$$E[Y] = \mu + \sigma\lambda\sqrt{\frac{2}{(1+\lambda^2)\pi}} \text{ e } Var[Y] = \sigma^2\left(1 - \frac{2\lambda^2}{(1+\lambda^2)\pi}\right). \tag{2.4}$$

Note que quando λ for 0, temos os mesmos resultados já conhecidos da distribuição normal.

2.2.2 Skew-t

A construção de famílias paramétricas de distribuições assimétricas que sejam analiticamente tratáveis e que possam acomodar valores práticos de assimetria e curtose, incluindo a distribuição skew-normal (normal e t-Student, no contexto simétrico) como caso particular, pode ser útil para a modelagem de dados. A distribuição ST, desde sua introdução à literatura, vem recebendo muita atenção, seja de modo teórico ao se estudar suas propriedades ou em aplicações numéricas como modelo para o ajuste de dados. Pode-se dizer que um dos grande motivos para gerar tanto interesse se deve ao seu elevado grau de flexibilidade, uma vez que seus parâmetros abrangem uma faixa admissível, com uma ampla variação das medidas associadas a assimetria e a curtose.

A distribuição skew-t usada é o caso univariado da proposta por Azzalini e Capitanio (2003), cuja função densidade de probabilidade é dada por

$$f(y|\mu,\sigma,\lambda,\tau) = \frac{2}{\sigma}f_{Z_1}(z)F_{Z_2}(\omega), y \in \mathbb{R}, \tag{2.5}$$

onde $-\infty < y < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \lambda < \infty$ e $\tau > 0$, f(.) é a função densidade de probabilidade de $Z_1 \sim TF(0,1,\tau)$, $F(.)$ é a função de distribuição

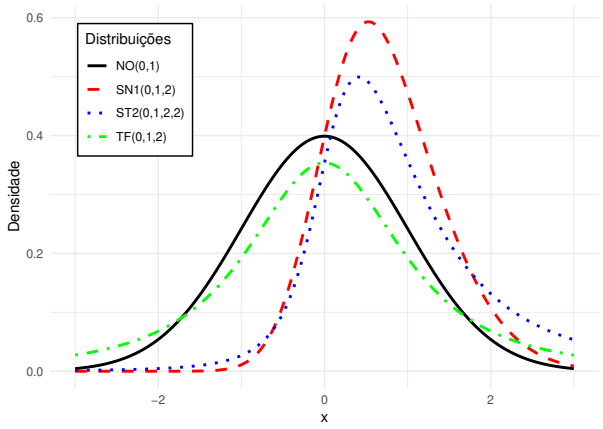


Figura 2: Densidades das distribuições

acumulada de $Z_2 \sim TF(0, 1, \tau + 1)$, sendo TF a distribuição t locação-escala, $z = \dfrac{y - \mu}{\sigma}$ e $\omega = \lambda \dfrac{(\tau + 1)}{(\tau + z^2)}$. Interessante notar que a distribuição t-student locação-escala é um caso particular da skew-t quando $\lambda = 0$. Note também que a distribuição ST permite combinar assimetria com as caudas pesadas e possui como casos particulares as distribuições skew-Cauchy (ν=1), skew-normal (ν tende ao infinito), t-Student (λ=0) e normal ($\lambda = 0$ e ν tende ao infinito).

Neste caso,

$$E[Y] = \mu + \sigma\Delta, \text{ para } \tau > 1 \text{ e } Var[Y] = \sigma^2\left(\frac{\tau}{\tau - 2} - \Delta^2\right) \text{ para } \tau > 2, \tag{2.6}$$

onde $\Delta = \dfrac{\lambda\sqrt{\tau}\Gamma(\frac{\tau-1}{2})}{\sqrt{1 + \lambda^2}\sqrt{\pi}}$.

Um comparativo das densidades dos modelos SN, ST, normal (NO) e t-Student (TF) pode ser visto na Figura 2. Observe que as distribuições normal e t-Student são simétricas, porém com o grau de liberdade igual a 2, temos as caudas pesadas. Com a introdução do parâmetro de assimetria igual a 2, temos tanto para a skew-normal quanto para a skew-t a presença de assimetria positiva, além de caudas pesadas no contexto da skew-t.

2.3 ALGORITMO RS

A fim de fazer a modelagem conjunta dos parâmetros de uma distribuição, o algoritmo de Rigby and Stasinopoulos (RS) será utilizado. As próximas seções são dedicadas a explicação de como o algoritmo iterativo de estimação funciona. Importante destacar que, no pacote gamlss, o algoritmo que estima modelos semi-paramétricos e não paramétricos tem o mesmo nome, mas funciona de forma diferente, e não é objetivo deste trabalho.

2.3.1 Funcionamento do algoritmo

Considere $y_1, \ldots, y_n$ independentes, tais que

$$Y \overset{\text{ind}}{\sim} \mathcal{D}(\mu_i, \sigma_i, \lambda_i, \tau_i)$$

e os parâmetros de locação, escala, assimetria e curtose podem ser modelados por variáveis conforme descrito em (2.2).

Os parâmetros do modelo paramétrico são estimados por máxima verossimilhança, conforme o logaritmo da função de verossimilhança que segue

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} log(f(y_i|\mu_i, \sigma_i, \lambda_i, \tau_i)) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta_i}), \tag{2.7}$$

onde $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ e f(.) podem ser as distribuições de interesse na Seção 2.2.

O algoritmo RS maximiza a log-verossimilhança para cada componente do parâmetro $\boldsymbol{\theta}$ por vez em um processo iterativo até a convergência, iniciando com $\boldsymbol{\mu}$, depois $\boldsymbol{\sigma}$, $\boldsymbol{\lambda}$ e finalizando com $\boldsymbol{\tau}$.

Esse algoritmo pode ser escrito como tendo duas partes:

i) as iterações externas e ii) as iterações internas.

Na etapa externa é verificada a convergência da deviance global ($-2\ell(\hat{\boldsymbol{\theta}})$), e, quando o valor alterar pouco (por padrão menos que 0.001) o algoritmo é finalizado. A etapa interna consiste em quatro subetapas que fornecem as estimativas atuais, calculam (ou atualizam) as variáveis e pesos associados a cada parâmetro que serão definidos na próxima subseção. Em cada subetapa, também é verificada a convergência de uma deviance. O critério padrão de finalização é uma diferença menor que 0.001, mas o pacote gamlss permite ajustar de forma independente os critérios de convergência das etapas externa e interna. Para obter novas estimativas dos parâmetros em cada subetapa, é realizada regressão por mínimos quadrados ponderados. Dessa regressão, são extraídos tanto os novos valores ajustados para os parâmetros quanto os coeficientes estimados. Para mais detalhes sobre a implementação, consulte o Apêndice A.

Nas próximas seções, serão apresentadas explicações adicionais sobre o funcionamento do algoritmo RS, especialmente no contexto do modelo sob distribuição SN. Uma das contribuições deste trabalho é fornecer um detalhamento do algoritmo RS aplicado a modelos assimétricos, com ênfase na distribuição skew-normal.

2.3.1.1 Definição das variáveis no contexto geral

Ao longo do algoritmo, algumas variáveis serão usadas repetidas vezes, são

- $\eta_i = g_k(fv_i)$: Relação entre preditor linear e o valor ajustado para cada parâmetro,
- $dr_i = \dfrac{d\theta_{k_i}}{d\eta_i}$: Derivada do parâmetro modelado θ_k em relação ao preditor linear η_i,
- $dldp_i = \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{k_i}}$: Derivada da função de log-verossimilhança em relação ao parâmetro θ_{k_i},
- $d2ldp2_i = \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_{k_i}^2}$: Segunda derivada da função de log-verossimilhança em relação ao parâmetro θ_{k_i},
- $wt_i = -(d2ldp2_i \times dr_i^2)$: Ponderação usada na atualização dos parâmetros,
- $wv_i = \eta_i + \dfrac{dldp_i\ dr_i}{wt_i}$: Variável de trabalho utilizada como variável resposta na regressão por mínimos quadrados ponderados,
- $d_i = -2\ell(\boldsymbol{\theta}_i)$: Deviance individual para a observação i,
- $dv_i = \sum_{i=1}^{n} d_i$: Deviance global somada para todas as observações,

onde η_i é o preditor linear, e θ_{k_i} o parâmetro sendo modelado.

2.3.2 Detalhes do Algoritmo RS no contexto do modelo SN

Nesta seção, vamos detalhar como o algoritmo RS funciona em cada etapa, no contexto assimétrico, em particular no modelo SN. Considere que

$$Y \overset{\text{ind}}{\sim} SN(\mu_i, \sigma_i^2, \lambda_i), i = 1, \ldots, n, \tag{2.8}$$

tal que,

$$\ell(\mu_i, \sigma_i^2, \lambda_i | y_i) = log(2) - log(\sigma_i) - \frac{log(2\pi)}{2} - \frac{(y_i - \mu_i)^2}{2\sigma_i^2} + log\left(\Phi\left(\lambda_i \frac{(y_i - \mu_i)}{\sigma_i}\right)\right). \tag{2.9}$$

Como trata-se de um algoritmo iterativo, é necessário atribuir valores iniciais arbitrários. Neste caso, vamos atribuir valores iniciais às estimativas de μ_i, σ_i e λ_i. Importante ressaltar que os valores iniciais são diferentes de acordo com a distribuição escolhida. A Tabela 1 indica quais são os valores iniciais para cada distribuição considerada nesse trabalho.

Com os valores iniciais atribuídos, damos início ao algoritmo começando pelo ciclo externo.

Tabela 1 – Valores iniciais de μ_i, σ_i, λ_i, τ_i para diferentes distribuições.

Distribuição	μ_i	σ_i	λ_i	τ_i
NO	$\frac{y_i+\bar{y}}{2}$	s_y		
TF	$\frac{y_i+\bar{y}}{2}$	s_y		10
SN	$\frac{y_i+\bar{y}}{2}$	$\frac{s_y}{4}$	0.1	
ST	$\frac{y_i+\bar{y}}{2}$	$\frac{s_y}{4}$	0.1	5

(1) Em posse dos últimos valores ajustados aos parâmetros μ_i, σ_i e λ_i (que nesse primeiro momento são os valores iniciais descritos na Tabela 1), vamos calcular d_i, conforme mostra a equação (2.10). Somando os valores obtidos, temos o valor da variável dv, dada pela equação (2.11).

$$d_i = -2\ell(\boldsymbol{\theta_i}) = -2\left(log(2) - log(\sigma_i) - \frac{log(2\pi)}{2} - \frac{(y_i-\mu_i)^2}{2\sigma_i^2} + log\left(\Phi\left(\lambda_i\frac{(y_i-\mu_i)}{\sigma_i}\right)\right)\right) \tag{2.10}$$

$$dv = \sum_{i=1}^{n} d_i. \tag{2.11}$$

Tendo calculado a deviance global, o próximo passo é iniciar o ciclo interno, que começa com a modelagem do parâmetro de locação.

(2) Modelagem do parâmetro de locação.

No começo dessa etapa, vamos herdar os valores utilizados no ciclo externo. Aqui, trataremos todos os parâmetros como constantes, exceto o parâmetro μ_i. Com isso, calculamos os valores de dr_i, $dldp_i$ e $d2ldp2_i$, com os quais calculamos wv_i e wt_i. O método do cálculo de fv_i, dr_i, $dldp_i$ e $d2ldp2_i$ pode variar de acordo com a distribuição escolhida, e estão exemplificadas abaixo para o caso skew normal.

$$\eta_i = fv_i \text{ (que vale } \tfrac{y_i+\bar{y}}{2} \text{ para a primeira interação)},$$
$$dr_i = 1, \text{ pois } \frac{d\mu}{d\eta} = \frac{d\eta}{d\eta} = 1,$$
$$dldp_i = \frac{\partial\ell(\theta_i)}{\partial\mu_i} = \frac{1}{\sigma_i^2}(y_i-\mu_i) - \frac{\lambda_i}{\sigma_i}\frac{\phi(\lambda_i\frac{(y_i-\mu_i)}{\sigma_i})}{\Phi(\lambda_i\frac{(y_i-\mu_i)}{\sigma_i})}\bigg|_{\mu_i=fv_i} = \frac{1}{\sigma_i^2}(y_i-fv_i) - \frac{\lambda_i}{\sigma_i}\frac{\phi\left(\lambda_i\frac{(y_i-fv_i)}{\sigma_i}\right)}{\Phi\left(\lambda_i\frac{(y_i-fv_i)}{\sigma_i}\right)},$$
$$d2ldp2_i = -(dldp_i)^2,$$
$$wt_i = -(d2ldp2_i \times dr_i^2) \text{ e}$$
$$wv_i = \eta_i + \frac{dldp_i\ dr_i}{wt_i}.$$

Com essas informações, é feita uma regressão linear por mínimos quadrados ponderados. As variáveis explicativas são as escolhidas para modelar o parâmetro de locação

(ou seja, x_i's), a variável resposta é a variável de trabalho wv_i e os pesos os valores de **wt**. Os valores ajustados nessa regressão (isto é, $\hat{y}_i$) são os novos valores de η_i, que por sua vez são os novos valores ajustados para μ_i. Com essas novas estimativas, podemos calcular a variável d_i seguindo a expressão descrita em (2.10), que ao ser somada nos gera uma nova deviance dv dada em (2.11). Se o valor de dv calculado for próximo ao valor de dv obtido na iteração anterior (por padrão utiliza-se uma diferença em valor absoluto menor que 10^{-3}), esta etapa é concluída e o algoritmo avança para a etapa (3). Caso contrário, volta-se ao início da etapa (2), utilizando os valores de μ_i ajustados nessa iteração como novos valores iniciais.

(3) Modelagem do parâmetro de escala.

A ideia é a mesma da etapa anterior, sendo herdado todos os últimos valores ajustados para μ_i, σ_i e λ_i. Contudo, agora apenas σ_i não será tratado como constante. Dessa forma, calculam-se novos valores de dr_i, $dldp_i$, $d2ldp2_i$, que por sua vez serão usados nos cálculos dos pesos wt_i e das variáveis de trabalho wv_i. A forma como são calculadas está explicitada abaixo. Importante notar que, ao contrário da etapa anterior, fv_i não adota simplesmente o valor de η_i, visto que, na modelagem do parâmetro de escala, a função de ligação escolhida não foi a identidade, mas sim a logarítmica. Por esse mesmo motivo, dr_i também foi calculado de forma diferente. Além disso, como $dldp_i$ agora é derivada de $\ell(\theta)$ com respeito a σ_i, uma nova equação foi geradada como resultado, conforme mostrado abaixo

$$\eta_i = log(fv_i),$$
$$dr_i = exp(\eta_i), \text{ pois } \frac{d\sigma}{d\eta} = \frac{dexp\{\eta\}}{d\eta} = exp\{\eta\},$$
$$dldp_i = \left.\frac{\partial \ell(\theta)}{\partial \sigma_i}\right|_{\sigma_i = fv_i} = \frac{-1}{fv_i} + \frac{1}{fv_i^3}(y_i - \mu_i)^2 - \frac{1}{fv_i^2}\lambda_i(y_i - \mu_i)\frac{\phi(\lambda_i \frac{(y_i-\mu_i)}{fv_i})}{\Phi(\lambda_i \frac{(y_i-\mu_i)}{fv_i})},$$
$$d2ldp2_i = -dldp_i^2,$$
$$wt_i = -(d2ldp2_i \times dr_i^2) \text{ e}$$
$$wv_i = \eta_i + \frac{dldp\ dr_i}{wt_i}.$$

Assim como na etapa (2), é feita uma regressão linear por mínimos quadrados ponderados. A diferença é que nesta regressão as variáveis explicativas são as escolhidas para modelar a variância (neste caso, z_i's). Os valores ajustados nessa nova regressão (ou seja, os valores de $\hat{y}_i$ obtidos na regressão por mínimos quadrados ponderados) são os valores de $\log(\sigma_i^2)$. Dessa forma, se aplicarmos a função exponencial aos valores de $\hat{y}_i$, temos novos valores ajustados para σ_i, com os quais podemos calcular a variável d_i, que ao ser somada considerando toda a amostra nos gera uma nova deviance dv. Analogamente ao feito na etapa anterior, comparamos essa nova deviance com a calculada baseada nas estimativas vigentes no começo dessa etapa. Caso a diferença seja menor que 10^{-3}, o algoritmo avança para a próxima etapa. Caso contrário, retorna-se ao início de (3),

utilizando os valores de σ_i ajustados nessa iteração como novos valores iniciais.

(4) Modelagem do parâmetro de assimetria.

Esta etapa é similar às anteriores, em especial à etapa (2), já que a função de ligação utilizada também é a identidade, logo, $dr_i = 1$. As únicas diferenças seriam que aqui trataríamos como constantes μ_i e σ_i e a variável $dldp_i$ seria referente à derivada da log-verossimilhança com respeito a λ_i, como mostrado abaixo. Além disso, as cováriaveis utilizadas na etapa de mínimos quadrados seriam, potencialmente, outras (w_i's).

$$\eta_i = fv_i \text{ herdando os últimos valores adotados na iteração anterior,}$$
$$dr_i = 1,$$
$$dldp_i = \left.\frac{\partial \ell(\theta)}{\partial \lambda_i}\right|_{\lambda_i = fv_i} = (\frac{y_i - \mu_i}{\sigma_i})\frac{\phi(fv_i\frac{(y_i-\mu_i)}{\sigma_i})}{\Phi(fv_i\frac{(y_i-\mu_i)}{\sigma_i})},$$
$$d2ldp2_i = -dldp_i^2,$$
$$wt_i = -(d2ldp2_i \times dr_i^2) \text{ e}$$
$$wv_i = \eta_i + \frac{dldp_i\ dr_i}{wt_i}.$$

Após os novos ajustes na regressão por mínimos quadrados ponderados, recalcula-se a deviance global para verificar o critério de convergência. Se o critério for atendido, o algoritmo avança para a etapa (5); caso contrário, retorna ao início da etapa (4).

(5) Recalculando a Global Deviance

Agora que todos os parâmetros de interesse já foram modelados, podemos calcular o critério que dirá se o algoritmo RS convergiu ou não. Dessa forma, vamos comparar o valor de dv mais recente (neste caso, o último valor de dv calculado na modelagem do parâmetro de assimetria) com o valor de dv calculado antes de se iniciar a modelagem do parâmetro de locação. Caso a diferença seja pequena (por padrão do pacote gamlss, a diferença deve ser menor que 0.001), o algoritmo é finalizado. Caso contrário, retorna-se à etapa (1).

Se a convergência for verificada, os valores de $\hat{\mu}_i$ serão os últimos ajustados na etapa (2) e, de forma análoga, os valores de $\hat{\sigma}_i$ serão os últimos ajustados na etapa (3), bem como os valores de $\hat{\lambda}_i$ serão os últimos ajustados na etapa (4). Além disso, os coeficientes estimados também serão os valores das últimas regressões por mínimos quadrados ponderados de cada etapa. Ou seja, os valores de $\hat{\boldsymbol{\beta}}$ obtidos nas últimas regressões serão os coeficientes estimados pelo algoritmo RS.

2.3.3 Exemplo prático do funcionamento do algoritmo RS no contexto da distribuição SN

Vamos apresentar um exemplo do funcionamento do algoritmo RS utilizando uma amostra de tamanho 3. Essa amostra será simulada de acordo com o modelo descrito

na equação (2.12). No nosso exemplo, definimos um critério de parada para o ciclo externo de 0.1 e para o ciclo interno de 0.5. A seguir, descreveremos os valores das variáveis relevantes $\mathbf{y}$ = (-0.3738165, 0.7478147, -0.8032220)^T, $\mathbf{x}$ = (-0.4689827, -0.2557522, 0.1457067)^T e $\mathbf{z} = (-0.6302355, 0.4047481, 0.1466527)^T$. Dessa forma, podemos observar como o algoritmo opera em cada etapa com base nessas definições. Note que sem perda de generalidade, nesta seção, consideramos apenas a modelagem dos parâmetros de locação e escala, considerando o parâmetro de assimetria fixo e conhecido, tal que $\lambda = -0.5$, onde

$$\begin{cases} y_i \sim SN(\mu_i, \sigma_i^2, \lambda), \\ \mu_i = 0.5x_i, \\ \log(\sigma_i^2) = -0.5z_i, \\ i = 1,\ 2, 3. \end{cases} \tag{2.12}$$

Para facilitar a compreensão, vamos explicar a notação que utilizaremos. Colocaremos índices acima de cada variável para indicar a qual etapa do algoritmo elas pertencem. O primeiro índice representa a iteração do ciclo externo, enquanto o segundo índice indica a etapa do clico interno: 1 para a modelagem do parâmetro $\boldsymbol{\mu}$ e 2 para modelagem do parâmetro $\boldsymbol{\sigma}$. O terceiro índice se refere à iteração específica dentro do ciclo interno. Se a variável não estiver em nenhuma etapa do ciclo interno, os dois últimos índeces serão zero. Por exemplo, o valor da variável dr que está na segunda etapa do ciclo externo, corresponde à modelagem do parâmetro $\boldsymbol{\mu}$ e foi calculado na terceira iteração do ciclo interno, terá como notação dr^{2,1,3}. Analogamente, a deviance global da quinta interação do ciclo externo será dada pela notação dv^{5,0,0}.

Os valores iniciais para $\boldsymbol{\mu}$ e $\boldsymbol{\sigma}$ são como definidos seguindo a Tabela 1, que, neste caso:
$\boldsymbol{\mu}^{1,1,0} = (-0.2584455, 0.3023701, -0.4731483)^T$,
$\boldsymbol{\sigma}^{1,1,0} = (0.2002124, 2002124, 2002124)^T$,
$-2\ell(\boldsymbol{\theta})^{1,1,0} = (-1.4555507, 6.2199973, 0.4112994)^T$ e
dv$^{1,1,0} = 5.175746$.

Em posse desses valores, entra-se no ciclo interno pela primeira vez, onde são calculadas as seguintes variáveis definidas na Seção 2.3.1.1,
$\boldsymbol{fv}^{1,1,1} = \boldsymbol{\mu}^{1,0,0} = (-0.2584455, 0.3023701, -0.4731483)^T$,
$\boldsymbol{dr}^{1,1,1} = (1, 1, 1)^T$,
$\boldsymbol{\sigma}^{1,1,1} = (0.2002124)^T$,
$-2\ell(\boldsymbol{\theta})^{1,1,1} = (-1.4555507, 6.2199973, 0.4112994)^T$,
dv$^{1,1,1} = 5.175746$,
$\boldsymbol{dldp}^{1,1,1} = (-1.319904, 15.147970, -7.342253)^T$,
$\boldsymbol{d2ldp2}^{1,1,1} = (1.742147, 229.460995, 53.908679)^T$,

$\boldsymbol{wt}^{1,1,1} = (1.742146, 229.460981, 53.908674)^T$ e
$\boldsymbol{wv}^{1,1,1} = (-1.016077, 0.368385, -0.609346)^T$.

A partir disso, realizamos uma regressão por mínimos quadrados ponderados. Na nossa análise, a variável resposta é $\boldsymbol{wv}^{1,1,1}$ enquanto a variável explicativa são os valores de $\mathbf{x}$, e os pesos correspondem aos valores de $\boldsymbol{wt}^{1,1,1}$. Os valores ajustados obtidos dessa regressão foram $\hat{\boldsymbol{y}} = (0.7253137, 0.3955382, -0.2253454)^T$. Esses valores representam as novas estimativas de $\boldsymbol{\mu}$. Com essas estimativas atualizadas, podemos prosseguir e recalcular a deviance global. Vamos comparar o novo valor da deviance global com o anterior, denotado como $\mathrm{dv}^{1,1,1}$. Se a diferença em valor absoluto for maior que 0.5, isso indica que precisamos repetir os cálculos das variáveis acima. Em outras palavras, reiniciaremos o ciclo interno para realizar uma nova regressão. Dessa forma, temos que
$\boldsymbol{\mu}^{1,1,2} = \boldsymbol{fv}^{1,2} = (0.7253137, 0.3955382, -0.2253454)^T$,
$\boldsymbol{dr}^{1,1,2} = (1, 1, 1)^T$, $\boldsymbol{\sigma}^{1,1,2} = (0.2002124, 0.2002124, 0.2002124)^T$,
$-2\ell(\boldsymbol{\theta})^{1,1,2} = (27.379024, 3.657498, 5.720475)^T$ e
$\mathrm{dv}^{1,1,2} = 36.757$.

Como a diferença entre $\mathrm{dv}^{1,1,2}$ e $\mathrm{dv}^{1,1,1}$ foi maior que 0.5, as variáveis são recalculadas, e é feita uma nova regressão, como mostrado a seguir:
$\boldsymbol{dldp}^{1,1,2} = (-27.396908, 12.358682, -14.036299)^T$,
$\boldsymbol{d2ldp2}^{1,1,2} = (-750.590555, -152.737032, -197.017699)^T$,
$\boldsymbol{wt}^{1,1,2} = (750.590555, 152.737032, 197.017699)^T$ e
$\boldsymbol{wv}^{1,1,2} = (0.688814, 0.476453, -0.296590)^T$.

Com os valores de $\hat{\boldsymbol{y}}$ obtidos, atualizamos as estimativas de $\boldsymbol{\mu}$, conforme visto abaixo
$\boldsymbol{\mu}^{1,1,3} = \boldsymbol{fv}^{1,3} = (0.7053196, 0.3846348, -0.2191335)^T$,
$\boldsymbol{dr}^{1,1,3} = (1, 1, 1)^T$,
$\boldsymbol{\sigma}^{1,1,3} = (0.2002124, 0.2002124, 0.2002124)^T$,
$-2\ell(\boldsymbol{\theta})^{1,1,3} = (26.293511, 3.930551, 5.895880)^T$ e
$\mathrm{dv}^{1,1,3} = 36.1199$.

Notamos que a diferença dessa deviance global para a anterior ($\mathrm{dv}^{1,1,2}$) é maior que 0.5, logo, devemos repetir o cálculo das derivadas, pesos, variável de trabalho e refazer a regressão.

Repetindo esse processo algumas vezes, obtemos:
$\mathrm{dv}^{1,1,5} = 34.4367$
$\boldsymbol{wt}^{1,1,6} = (646.932193, 185.790021, 215.398727)^T$
$\boldsymbol{wv}^{1,1,6} = (0.607979, 0.426357, -0.269242)^T$.

Os valores ajustados dessa regressão foram $\hat{\boldsymbol{y}} = (0.628702, 0.342853, -0.195329)^T$, e temos que $\mathrm{dv}^{1,1,6} = 33.9496$.

Comparando com a deviance global anterior ($\text{dv}^{1,1,5}$), constamos que a diferença é menor que 0.5, e, portanto, finalizamos essa etapa do ciclo interno. Inicia-se agora a modelagem do parâmetro $\boldsymbol{\sigma}$, tal que
$\boldsymbol{fv}^{1,2,1} = \boldsymbol{\sigma}^{1,1,6} = (0.2002124, 0.2002124, 0.2002124)^T$,
$\boldsymbol{dr}^{1,2,1} = (0.2002124, 0.2002124, 0.2002124)^T$,
$\boldsymbol{\mu}^{1,2,1} = \boldsymbol{\mu}^{1,1,6} = (0.628702, 0.342853, -0.195329)^T$,
$-2\ell(\boldsymbol{\theta})^{1,2,1} = (22.319971, 5.042700, 6.586903)^T$,
$\text{dv}^{1,2,1} = 33.9496$,
$\boldsymbol{dldp}^{1,2,1} = (120.017824, 23.189190, 40.028658)^T$,
$\boldsymbol{d2ldp2}^{1,2,1} = (-14404.277987, -537.738548, -1602.293423)^T$,
$\boldsymbol{wt}^{1,2,1} = (577.395509, 21.555251, 64.227935)^T$ e
$\boldsymbol{wv}^{1,2,1} = (-1.566760, -1.392988, -1.483599)^T$.

Agora, realizamos a regressão utilizando o método de mínimos quadrados ponderados. Os coeficientes ajustados obtidos com essa regressão são -1.4636097, 0.9399553 e 0.3405747. Em seguida, aplicamos a função exponencial a esses valores para calcular os novos valores ajustastados de $\boldsymbol{\sigma}$, ou seja,
$\boldsymbol{\sigma}^{1,2,2} = (0.231399, 2.559868, 1.405755)^T$,
$\boldsymbol{dr}^{1,2,2} = (0.231399, 2.559868, 1.405755)^T$ e
$\boldsymbol{\mu}^{1,2,2} = (0.628702, 0.342853, -0.195329)^T$.

Assim, obtemos que
$\boldsymbol{dldp}^{1,2,2} = (76.429884, -0.354638, -0.680702)^T$,
$\boldsymbol{d2ldp2}^{1,2,2} = (-5841.527218, -0.125768, -0.463355)^T$,
$\boldsymbol{wt}^{1,2,2} = (312.788626, 0.824149, 0.915658)^T$,
$\boldsymbol{wv}^{1,2,2} = (-1.407068, -0.161577, -0.704466)^T$,
$-2\ell(\boldsymbol{\theta})^{1,2,2} = (16.324780, 3.873056, 2.389999)^T$ e
$\text{dv}^{1,2,2} = 22.58784$.

Como a deviance global atual é muito diferente da deviance global anterior, é necessário refazer os cálculos das estimativas de $\boldsymbol{\sigma}$ utilizando o método de mínimos quadrados ponderados. A partir dessa nova regressão, os valores ajustados são $\hat{\boldsymbol{y}} = (-1.4045657, 0.9020363, 0.3268355)^T$. Ao aplicar a função exponencial, as novas estimativas de são 0.2454736, 2.4646167 e 1.3865734. Com essas novas estimativas, recalculam-se os pesos, a variável de trabalho e a deviance global para verificar a convergência do modelo. Esse processo é repetido algumas vezes, tal que $\text{dv}^{1,2,11} = 10.6187$, $\text{dv}^{1,2,12} = 8.0208$ e $\text{dv}^{1,2,13} = 6.3186$.

Na etapa dos mínimos quadrados, obtemos $\hat{\boldsymbol{y}} = (0.0032704176, -0.0021003184, -0.0007610099)^T$ que aplicado na função exponencial nos traz as estimativas para $\boldsymbol{\sigma}^{1,2,14}$ como 1.00327577120197, 0.997901885688574 e 0.999239279554774. Calculando a deviance global, temos que $\text{dv}^{1,2,14} = 6.3217$. Essa deviance comparada com a anterior ($\text{dv}^{1,2,13}$)

tem uma diferença menor que 0.5, e, portanto, essa etapa é finalizada.

Como λ foi considerado um parâmetro fixo, não será modelado. Dessa forma, o ciclo interno se encerra por completo. Agora, vamos comparar essa última deviance global calculada ($\text{dv}^{1,2,14} = 6.3217$) com a deviance global inicial (ou seja, $\text{dv}^{1,0,0} = 5.175746$). Como a diferença é maior que 0.1, repete-se o ciclo externo, mas agora os valores iniciais são os últimos ajustados aos parâmetros $\boldsymbol{\mu}$ ($(0.628702, 0.342853, -0.195329)^T$) e $\boldsymbol{\sigma}$ ($(1.0032767, 0.997902, 0.999239)^T$). Além disso, temos que $\text{dv}^{2,0,0} = 6.3217$.

Agora, reiniciamos o ciclo interno começando com a modelagem do parâmetro $\boldsymbol{\mu}$. A deviance inicial calculada é $\text{dv}^{2,1,1} = 6.3217$. Após essa etapa, prosseguimos com o cálculo de:
$\boldsymbol{dr}^{2,1,1} = (1, 1, 1)^T$,
$\boldsymbol{dldp}^{2,1,1} = (-0.742137138389878, 0.87333940426722, -0.301155405960204)^T$ e
$\boldsymbol{d2ldp2}^{2,1,1} = (-0.550767532177517, -0.762721715045823, -0.0906945785390552)^T$.

Logo,
$\boldsymbol{wt}^{2,1,1} = (0.550767532177517, 0.762721715045823, 0.0906945785390552)^T$ e
$\boldsymbol{wv}^{2,1,1} = (-0.718757674349582, 1.48788288712527, -3.51587421173876)^T$.

A regressão feita com essas variáveis produz como resultado $\hat{\boldsymbol{y}} = (0.4095745, 0.2233549, -0.1272494)^T$ que são as novas estimativas para $\boldsymbol{\mu}$. Importante ressaltar que dessa regressão o coeficiente calculado foi de -0.8733254.

Agora, recalculamos a deviance global, obtendo $\text{dv}^{2,1,2} = 6.322191$. Comparando esse valor com o anterior ($\text{dv}^{2,1,1} = 6.3217$), observamos que a diferença é menor que 0.5. Com isso, concluímos a modelagem do parâmetro $\boldsymbol{\mu}$, e iniciamos a modelagem do parâmetro $\boldsymbol{\sigma}$. Assim, definimos $\text{dv}^{2,2,1} = 6.3217$, e seguimos com o cálculo de:
$\boldsymbol{dldp}^{2,2,1} = (-0.609697570997367, -0.469251423809317, -0.744441667126885)^T$ e
$\boldsymbol{d2ldp2}^{2,2,1} = (-0.371731128080089, -0.220196898747071, -0.554193395754656)^T$, e disso temos que
$\boldsymbol{wt}^{2,2,1} = (0.374170529255223, 0.219273871543057, 0.553350543970413)^T$ e
$\boldsymbol{wv}^{2,2,1} = (-1.63153177266851, -2.13763464993793, -1.34507224851378)^T$.

Realizando a regressão com essas variáveis, obtemos $\hat{\boldsymbol{y}} = (\ -0.27549586, 0.17692818, 0.06410652)^T$, com o coeficiente calculado de 0.4371316. Ao aplicar a função exponencial nesses resultados, as novas estimativas para $\boldsymbol{\sigma}$ são (0.759195577330785, 1.19354537061228, 1.06620596414822)^T. Com esses novos valores, recalculamos a deviance global, obtendo $\text{dv}^{2,2,2} = 6.3778$. Comparando esse valor com a última deviance global obtida durante a modelagem do $\boldsymbol{\sigma}$ ($\text{dv}^{2,2,1} = 6.3222$), temos que a diferença é menor que 0.5, e, portanto, a modelagem do $\boldsymbol{\sigma}$ está finalizada.

Por fim, vamos conferir se a deviance global convergiu no ciclo externo também, comparando $\text{dv}^{2,2,2} = 6.3778$ e $\text{dv}^{2,0,0} = 6.3217$. Como a diferença é menor que 0.1, é considerado que o ciclo externo convergiu, logo, o algoritmo é finalizado.

Vamos agora extrair as informações finais do ajuste. Os últimos valores ajustados para $\boldsymbol{\mu}$ são as novas estimativas para esse parâmetro $(0.4095745, 0.2233549, -0.1272494)$. Analogamente, os últimos valores ajustados para $\boldsymbol{\sigma}$ são as novas estimativas para esse parâmetro $(0.7591956, 1.1935454, 1.0662060)$. Por fim, os coeficientes estimados são herdados das regressões por mínimos quadrados ponderados, ou seja, $\hat{\beta}_1 = -0.8733254$ e $\hat{\gamma}_1 = 0.4371316$. É interessante notar que, nesta simulação, os verdadeiros valores de β_1 e γ_1 não foram bem recuperados. Isso ocorreu principalmente por dois motivos: primeiro, o tamanho da amostra é muito pequeno, o que naturalmente reduz a precisão das estimativas. Além disso, para tornar o exemplo mais didático, os critérios de convergência dos ciclos interno e externo foram ajustados para valores muito superiores ao recomendado (0.001), o que comprometeu ainda mais a recuperação dos valores verdadeiros dos parâmetros. No Capítulo 3, serão realizados estudos de simulação utilizando diferentes tamanhos de amostras e aplicando um critério de convergência mais adequado. Nesta seção, o foco foi apenas apresentar o funcionamento do algoritmo RS de maneira didática, explicando passo a passo como ele funciona com um exemplo ilustrativo.

2.4 CÁLCULO DO ERRO PADRÃO

No modelo GAMLSS paramétrico, o cálculo do erro padrão pode ser feito via bootstrap (seja ele paramétrico, não paramétrico ou dos resíduos) ou via inferência baseada nas propriedades assintóticas dos estimadores de máxima verossimilhança. Essa segunda opção será a utilizada neste trabalho.

Dessa forma, para o vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, \lambda, \tau)^T$, temos $\boldsymbol{\beta} = (\boldsymbol{\beta_\mu}, \boldsymbol{\beta_\sigma}, \boldsymbol{\beta_\lambda}, \boldsymbol{\beta_\tau})^T = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4})$, de dimensão p, $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, [I_F(\boldsymbol{\beta})]^{-1})$ é uma normal multivariada de dimensão p, onde $I_F(\boldsymbol{\beta})$ é a matriz de informação esperada de Fisher, definida como

$$I_F(\boldsymbol{\beta}) = E\left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right). \tag{2.13}$$

Entretanto, se não for possível obter de forma analítica a matriz de informação esperada de Fisher, utiliza-se a matriz de informação observada de Fisher, definida como

$$J(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}. \tag{2.14}$$

Na prática, $I_F(\boldsymbol{\beta})$ e $J(\boldsymbol{\beta})$ são usualmente desconhecidos e os substituímos por $I_F(\hat{\boldsymbol{\beta}})$ e $J(\hat{\boldsymbol{\beta}})$ que são as matrizes avaliadas na estimativa de máxima verossimilhança do $\boldsymbol{\beta}$.

A raiz quadrada da diagonal principal da matriz inversa de $J(\hat{\boldsymbol{\beta}})$ é o erro padrão de $\hat{\boldsymbol{\beta}}$. Assim, assumindo um nível de significância de 5%, podemos calcular intervalos de confiança para qualquer coeficiente m pertencente a $\boldsymbol{\beta_k}$ como

$$\hat{\beta}_{k_m} \pm 1.96 se(\hat{\beta}_{k_m}), \tag{2.15}$$

onde $se(\hat{\beta}_{k_m})$ é o erro padrão do coeficiente estimado $\hat{\beta}_{k_m}$.

Ao interpretarmos esses intervalos de confiança, podemos realizar testes de hipótese para verificar a significância dos coeficientes de regressão no modelo. Um parâmetro será considerado não significativo ao nível de confiança de 95% se o intervalo de confiança para o seu estimador incluir o valor 0. Caso contrário, o parâmetro é considerado significativo, pois há evidências suficientes para rejeitar a hipótese nula de que ele seja zero.

2.5 DETERMINAÇÃO DAS FUNÇÕES DE LIGAÇÃO

A função de ligação adequada depende do suporte dos parâmetros da distribuição da variável de resposta. Neste trabalho, consideramos as distribuições SN (e sua versão simétrica, normal) e ST (e sua versão simétrica, t-Student). Portanto, consideramos as funções de ligação padrão do próprio pacote gamlss para cada uma das distribuições estudadas, ou seja, as funções de ligação padrão são a identidade para os parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\lambda}$ e logaritmo natural para os parâmetros $\boldsymbol{\sigma}$ e $\boldsymbol{\tau}$.

2.6 CRITÉRIO DE SELEÇÃO DE VARIÁVEIS

Dada uma distribuição para a variável resposta com um conjunto de parâmetros $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\lambda}$ e $\boldsymbol{\tau}$, a significância estatística pode ser empregada para encontrar os melhores preditores para um dado parâmetro da distribuição da variável de resposta. Como critério, adotamos o nível de significância de 5%, e sempre será removida a variável não significativa com maior p-valor, independentemente de qual parâmetro esteja modelando. Após remover uma variável, o modelo é reajustado, e repete-se o processo até que o modelo final contenha apenas preditores que estejam relacionados à variável resposta.

2.7 CRITÉRIO DE SELEÇÃO DE MODELOS

Para este trabalho, tendo um modelo final para cada distribuição estudada, devemos selecionar qual dos modelos se ajustou melhor aos dados. Uma das métricas possíveis para a seleção do modelo é o AIC (Akaike information criterion), definido como

$$AIC = -2\ell(\boldsymbol{\theta}) + 2p, \tag{2.16}$$

onde p é a quantidade de parâmetros estimados do modelo e $\ell(\boldsymbol{\theta})$ é a log-verossimilhança dos dados observados.

Adicionalmente, podemos usar também o BIC (Schwarz Bayesian criterion), definido como

$$BIC = -2\ell(\boldsymbol{\theta}) + log(n)p. \tag{2.17}$$

Note que, para grandes amostras, o BIC penaliza mais a adição de novos parâmetros. Quanto menor o AIC ou o BIC, melhor o ajuste do modelo aos dados. Dessa forma, conseguimos comparar modelos, mesmo que não sejam aninhados. Caso o AIC indique um modelo como melhor, mas o BIC outro, e não houver motivo para dar preferência a um critério em particular, devemos complementar com a análise de resíduos, tema da próxima seção.

2.8 ANÁLISE DOS RESÍDUOS

Depois de ajustar o modelo é importante analisar os resíduos para verificar tanto a qualidade do ajuste quanto a plausibilidade dos pressupostos do modelo.

Neste trabalho, utilizaremos os resíduos quantílicos normalizados introduzidos em Dunn e Smyth (1996). A principal vantagem dos resíduos quantílicos normalizados é que, qualquer que seja a distribuição da variável resposta, os resíduos sempre têm uma distribuição normal padrão quando o modelo assumido está correto. Como a verificação de suposições do modelo por meio da normalidade dos resíduos é bem estabelecida na literatura estatística, os resíduos quantílicos normalizados fornecem uma maneira familiar de verificar a adequação de um modelo ajustado.

Usualmente, cinco gráficos são utilizados para verificar o ajuste do modelo. O primeiro gráfico exibe os resíduos quantílicos normalizados contra os valores ajustados para μ. Nele, espera-se que nenhum padrão seja encontrado, indicando que os resíduos não ajudam a prever a variável resposta. O segundo gráfico mostra os valores dos resíduos contra o índice, onde também não se espera a presença de padrões, ou seja, os pontos devem estar aleatoriamente distribuídos. O terceiro gráfico apresenta a densidade kernel estimada para os resíduos quantílicos, a qual deve se aproximar da densidade da distribuição normal. O quarto gráfico é o QQ-plot dos resíduos quantílicos, no qual os pontos devem estar próximos da reta vermelha que indica o ajuste perfeito. Já o quinto gráfico, conhecido como worm plot (gráfico de "minhoca"), é uma modificação do QQ-plot dos resíduos que remove a tendência. Ele foi introduzido por Buuren e Fredriks (2001) e demonstra graficamente o quão bem o modelo se ajusta aos dados, além de indicar como o modelo pode ser melhorado. Se o modelo estiver bem especificado, os pontos do worm plot estarão próximos da linha horizontal, e 95% deles estarão entre as curvas pontilhadas superior e inferior, que correspondem às bandas de confiança de 95%, sem apresentar desvios sistemáticos. Além disso, é traçada uma curva em vermelho no gráfico. Essa curva é um ajuste aos dados do gráfico, feito por meio de um polinômio cúbico que ajuda a entender o formato da minhoca, veja na Tabela 2 os diferentes formatos da minhoca (ou curva ajustada).

A Figura 3 mostra como os resíduos devem se comportar se o ajuste for adequado, conforme descrito nos quatro primeiros gráficos. O worm plot ideal está apresentado na

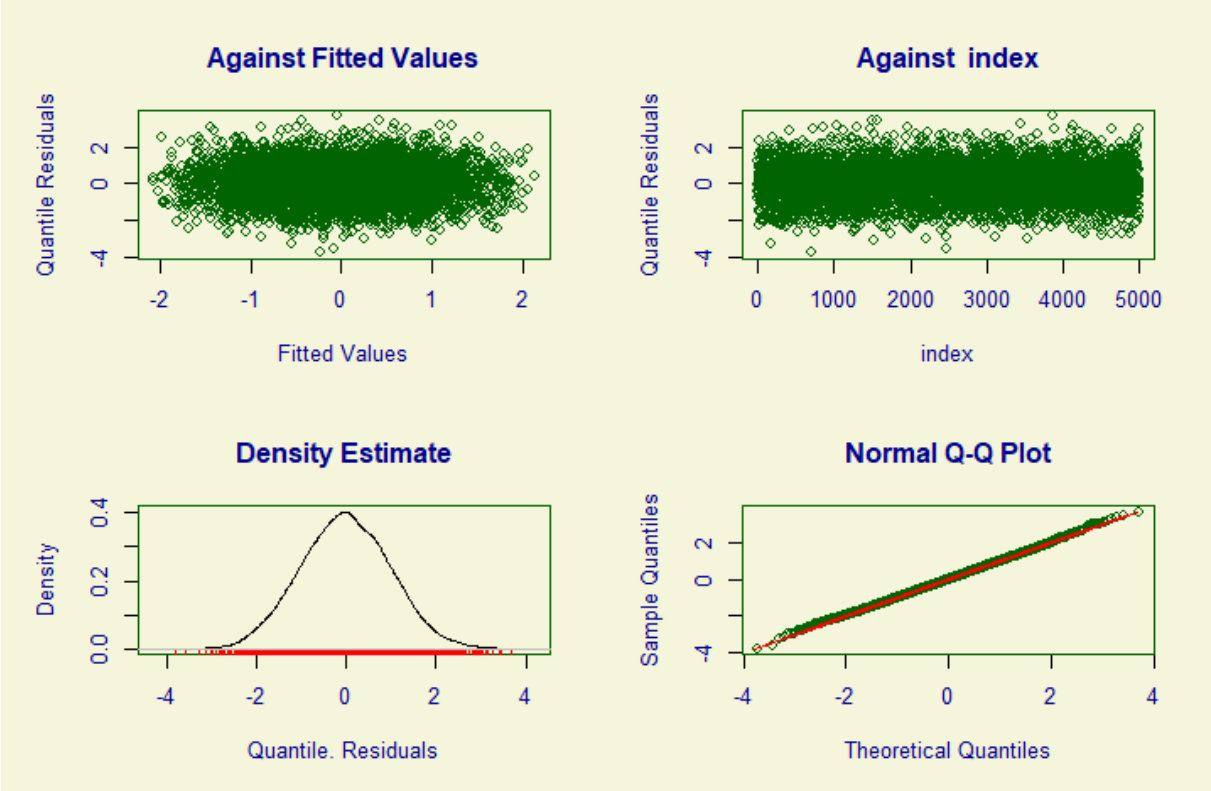


Figura 3: Gráficos ideais de resíduos.

Figura 4, e, se ampliarmos os limites do gráfico para valores maiores, vemos que o formato esperado dos resíduos se mantém sem desvios sistemáticos, conforme exibido na Figura 5. A Tabela 2 nos auxilia na interpretação dos formatos que a curva ajustada pode assumir e como isso ajuda a identificar maneiras de melhorar o ajuste do modelo. Esses formatos são demonstrados na Figura 6, retirada de Rigby et al. (2019).

É importante destacar que, ao analisar o worm plot, devemos ter cautela. Observando a Figura 4 e seguindo as recomendações da Tabela 2, poderíamos interpretar equivocadamente que os resíduos formam um padrão em U, sugerindo que a assimetria ajustada está abaixo do ideal. No entanto, isso não é verdade. Esse erro decorre do efeito de escala, pois, como o modelo foi muito bem ajustado, os resíduos são pequenos (a escala varia de -0.15 a 0.15). Quando ampliamos os limites do gráfico para -1.5 a 1.5, conforme mostrado na Figura 5, o formato em U desaparece. Os códigos para a confecção desses gráficos estão disponíveis no Apêndice D.

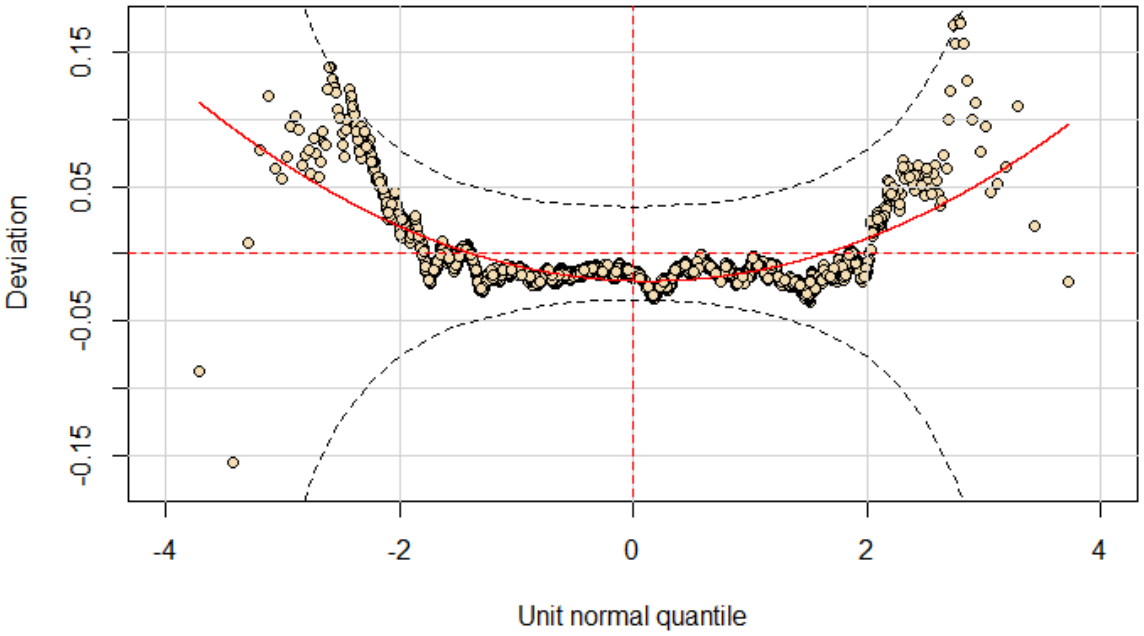


Figura 4: Worm plot ideal.

Forma da minhoca (ou da curva ajustada)	Resíduos	Distribuição ajustada
Acima da origem	Média muito alta	Média ajustada muito baixa
Abaixo da origem	Média muito baixa	Média ajustada muito alta
Inclinação para cima	Variância muito alta	Variância ajustada muito baixa
Inclinação para baixo	Variância muito baixa	Variância ajustada muito alta
Formato de U	Assimetria positiva	Assimetria ajustada muito baixa
Formato de U invertido	Assimetria negativa	Assimetria ajustada muito alta
Formato de S começando por baixo	Leptocurtose	Peso das caudas da distribuição utilizada muito baixo
Formato de S começando por cima	Platicurtose	Peso das caudas da distribuição utilizada muito alto

Tabela 2 – Indicativos do que deve ser melhorado com base no worm plot.

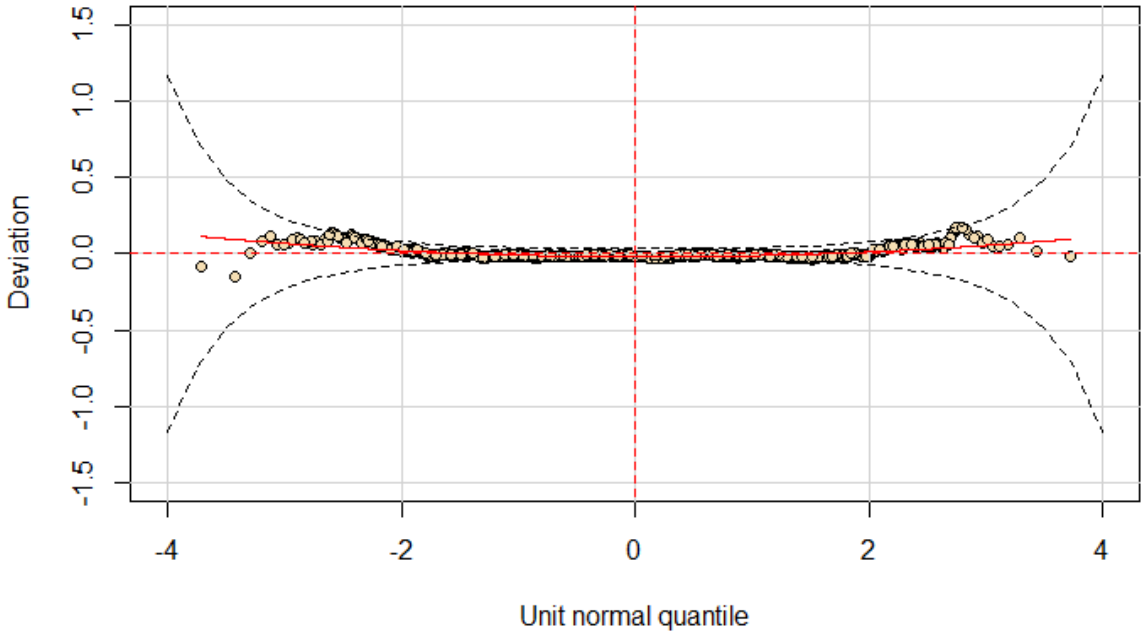


Figura 5: Worm plot com limites do eixo vertical ampliados.

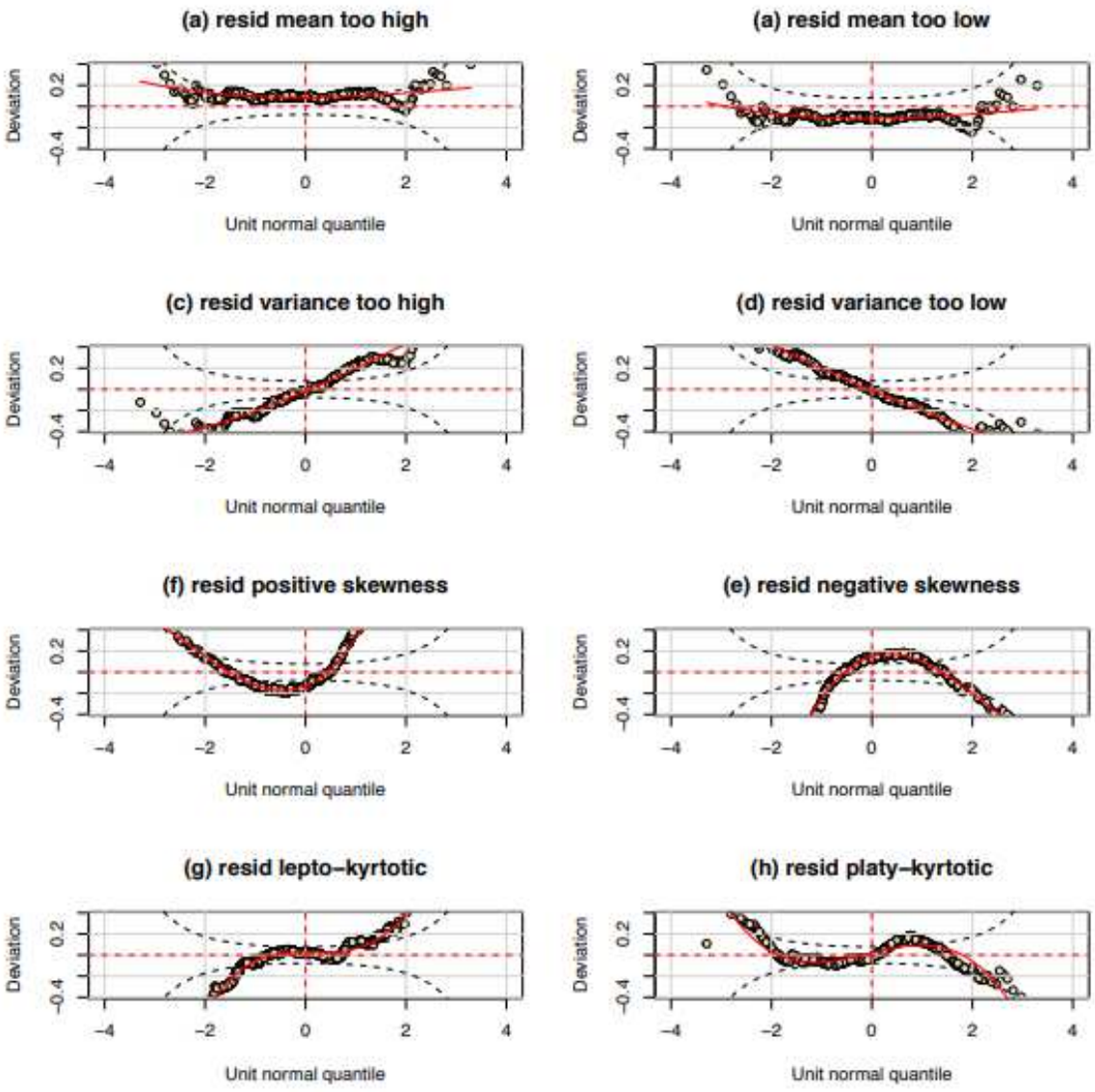


Figura 6: Problemas detectados via worm plot.

3 ESTUDO DE SIMULAÇÃO

3.1 AMBIENTE DE DESENVOLVIMENTO

Os hardwares e softwares utilizados foram:

- **Sistema Operacional:** Windows 10 (versão 22H2).
- **Especificações do dispositivo:**
  - Processador: Intel Core 17 6500U CPU @ 2.50GHz.
  - Memória RAM: 8GB DDR4-2133MHz.
  - Placa gráfica: NVIDIA NVIDIA GeForce 920MX.
- **Linguagem de Programação:** R versão 4.3.0 (R Core Team, 2023).
- **Ambiente de Desenvolvimento Integrado:** RStudio versão 2023.6.2.561 (Posit team, 2023).
- **Pacote utilizado:**
  - gamlss - Rigby e Stasinopoulos (2005) - versão 4.3.3.

3.2 MODELAGEM CONJUNTA DOS PARÂMETROS DO MODELO SN

O principal objetivo deste estudo é avaliar o desempenho do algoritmo RS na obtenção de estimativas de máxima verossimilhança do GAMLSS sob distribuição SN, e examinar o comportamento desse modelo sob diferentes configurações, comparando-o, quando possível, com outros algoritmos de estimação.

3.2.1 Modelagem conjunta dos parâmetros de locação e escala do modelo SN

Nesta seção, usamos simulações de Monte Carlo para avaliar o desempenho dos estimadores de máxima verossimilhança dos parâmetros associados com a locação e escala do GAMLSS sob distribuição SN. O estudo de simulação foi projetado para observar mudanças (alterações) nas estimativas variando os tamanhos amostrais.

Seguindo Li e Wu (2014), os dados foram artificialmente gerados provenientes de um caso particular do modelo definido por (2.1) e (2.2) e apresentado a seguir.

$$\begin{cases} y_i \sim SN(\mu_i, \sigma_i^2, \lambda), \\ \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \\ \log(\sigma_i^2) = \boldsymbol{z}_i^T \boldsymbol{\gamma}, \\ i = 1,\ 2, \ldots, n, \end{cases} \tag{3.1}$$

com os vetores de covariáveis x^T e z^T, cujas componentes foram geradas de uma distribuição uniforme (-1,1). Geramos 1000 conjuntos de dados provenientes do modelo (3.1) com a seguinte configuração: $\boldsymbol{\beta} = (1, 0.7, 0.5)^T$, $\boldsymbol{\gamma} = (-0.5, 0.3, 0.2)^T$ e três cenários (valores diferentes) para λ=-0.5, 0 ou 0.5. Diferentes tamanhos amostrais, isto é, n = 50, 100 e 150 foram considerados. O objetivo destes cenários é mostrar o comportamento dos estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ obtidos via algoritmo RS quando λ é considerado fixo e conhecido. Nestes cenários, Li e Wu (2014) propuseram estimar simultaneamente os coeficientes das equações de μ e σ por máxima verossimilhança usando o método iterativo de Newton Raphson e o desempenho dos estimadores $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\gamma}}$ será avaliado usando o erro quadrático médio (EQM), definidos como $EQM(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ e $EQM(\hat{\boldsymbol{\gamma}}) = E(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$, respectivamente. Na prática, o valor esperado foi substituído pela média do quadrado da diferença entre o valor verdadeiro e o estimado dos coeficientes.

O código utilizado para simular uma amostra para n=50 e $\lambda = -0.5$ está apresentado no Apêndice A, enquanto a simulação de Monte Carlo completa com o cálculo do EQM no Apêndice B.

As Tabelas 3, 4 e 5 apresentam o EQM e a média das estimativas de Monte Carlo dos parâmetros, obtidas via algoritmo RS, em todas as amostras do GAMLSS sob distribuição SN nos diferentes cenários estudados. Para fins de comparação, a Tabela 6 apresenta os resultados das simulações para $\lambda = -0.5$ (cenário 1), mas considerando a estimação de máxima verossimilhança proposta em Li e Wu (2014), que utilizou o algoritmo de Newton-Raphson..

Tabela 3 – Cenário 1 - $\lambda = -0.5$ - EQM e média das estimativas de Monte Carlo dos parâmetros do GAMLSS sob distribuição SN.

Parâmetro	Valor Verdadeiro	Média das Estimativas			EQM das Estimativas		
		n = 50	n = 100	n = 150	n = 50	n = 100	n = 150
β_1	1	0.9982	1.0043	0.9946			
β_2	0.7	0.7067	0.7011	0.6976	0.1529	0.0689	0.0427
β_3	0.5	0.5090	0.5042	0.5058			
γ_1	-0.5	-0.5394	-0.5131	-0.5129			
γ_2	0.3	0.3164	0.3141	0.3057	0.1472	0.0541	0.0319
γ_3	0.2	0.2137	0.2140	0.2014			

Analisando as Tabelas 3, 4 e 5, observamos que independente do valor adotado para λ, temos um padrão muito claro nos resultados. Quanto maior o tamanho da amostra, mais próximas dos valores reais são as médias das estimativas (ou seja, menor viés), além de ocorrer a diminuição dos valores de EQM, como esperado. Isso concorda essencialmente com as propriedades assintóticas do estimador de máxima verossimilhança, isto é, consistência. As estimativas obtidas via algoritmo RS podem ser comparadas às obtidas por Li e Wu (2014), via método Newton Raphson, conforme exemplificado na

Tabela 4 – Cenário 2 - $\lambda = 0$ - EQM e média das estimativas de Monte Carlo dos parâmetros do GAMLSS sob distribuição SN.

Parâmetro	Valor Verdadeiro	Média das Estimativas			EQM das Estimativas		
		n = 50	n = 100	n = 150	n = 50	n = 100	n = 150
β_1	1	1.0091	0.9991	1.0003			
β_2	0.7	0.6851	0.6956	0.6977	0.1790	0.0770	0.0500
β_3	0.5	0.4966	0.5042	0.5032			
γ_1	-0.5	-0.5207	-0.5162	-0.5142			
γ_2	0.3	0.3122	0.3107	0.3014	0.1631	0.0568	0.0351
γ_3	0.2	0.2108	0.2084	0.2093			

Tabela 5 – Cenário 3 - $\lambda = 0.5$ - EQM e média das estimativas de Monte Carlo dos parâmetros do GAMLSS sob distribuição SN.

Parâmetro	Valor Verdadeiro	Média das Estimativas			EQM das Estimativas		
		n = 50	n = 100	n = 150	n = 50	n = 100	n = 150
β_1	1	0.9967	1.0004	1.0037			
β_2	0.7	0.7055	0.7071	0.7053	0.1523	0.0679	0.0442
β_3	0.5	0.4951	0.4980	0.5025			
γ_1	-0.5	-0.5310	-0.5154	-0.5154			
γ_2	0.3	0.3208	0.3052	0.3049	0.1395	0.0553	0.0335
γ_3	0.2	0.2130	0.2097	0.2009			

Tabela 6 – Cenário 1 - $\lambda = -0.5$ - EQM e média das estimativas de Monte Carlo dos parâmetros do GAMLSS sob distribuição SN, obtidas em Li e Wu (2014).

Parâmetro	Valor Verdadeiro	Média das Estimativas			EQM das Estimativas		
		n = 50	n = 100	n = 150	n = 50	n = 100	n = 150
β_1	1	1.0061	1.0052	1.0042			
β_2	0.7	0.7113	0.6965	0.7003	0.1848	0.0810	0.0514
β_3	0.5	0.4924	0.4989	0.5000			
γ_1	-0.5	-0.5269	-0.5101	-0.5093			
γ_2	0.3	0.3219	0.3127	0.2989	0.5356	0.2163	0.1338
γ_3	0.2	0.2280	0.2103	0.2068			

Tabela 6. Pode-se notar que as conclusões feitas para os resultados obtidos via algoritmo RS também podem ser feitas para o algoritmo de Newton Raphson, porém o valor do EQM, diferiu bastante, principalmente em relação aos coeficientes associadas à escala. Então, podemos concluir que o algoritmo RS utiizado, neste trabalho, produz estimativas mais precisas.

3.2.2 Modelagem conjunta dos parâmetros de locação, escala e assimetria do modelo SN

Nesta seção, usamos simulações de Monte Carlo para avaliar o desempenho dos estimadores de máxima verossimilhança dos parâmetros associadas com a locação, escala e assimetria do GAMLSS sob distribuição SN. O estudo de simulação foi projetado para

observar mudanças (alterações) nas estimativas variando os tamanhos amostrais.

Os dados foram artificialmente gerados provenientes do modelo definido em (3.2) e apresentado a seguir.

$$\begin{cases} y_i \sim SN(\mu_i, \sigma_i^2, \lambda_i), \\ \hat{\mu}_i = x_{1i} + 0.75x_{2i} + 0.5x_{3i}, \\ log(\hat{\sigma}_i^2) = -0.5z_{1i} + 0.3z_{2i} + 0.2z_{3i}, \\ \hat{\lambda}_i = -0.2w_{1i} + 0.5w_{2i} + 1w_{3i}, \\ i = 1,\ 2, \ldots, n, \end{cases} \tag{3.2}$$

onde x_{ji}'s, z_{ji}'s e w_{ji}'s foram geradas de uma distribuição uniforme (-1,1). Geramos 1000 conjuntos de dados provenientes do modelo (3.2) e os valores adotados para os coeficientes de $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ e $\boldsymbol{\omega}$ foram: $\boldsymbol{\beta} = (1, 0.75, 0.5)^T$, $\boldsymbol{\gamma} = (-0.5, 0.3, 0.2)^T$ e $\boldsymbol{\omega} = (-0.2, 0.5, 1)^T$. Diferentes tamanhos amostrais (n = 50, 100 e 150) foram considerados.

A Tabela 7 apresenta o EQM e a média das estimativas de Monte Carlo dos parâmetros, obtidas via algoritmo RS, em todas as amostras do GAMLSS sob distribuição SN.

Tabela 7 – EQM e média das estimativas de Monte Carlo dos parâmetros do GAMLSS sob distribuição SN.

Parâmetro	Referência	Média			EQM		
		$n = 50$	$n = 100$	$n = 150$	$n = 50$	$n = 100$	$n = 150$
β_1	1.000	0.995	0.992	0.999			
β_2	0.750	0.763	0.751	0.750	0.165	0.065	0.042
β_3	0.500	0.508	0.500	0.501			
γ_1	-0.500	-0.528	-0.513	-0.514			
γ_2	0.300	0.325	0.310	0.312	0.152	0.051	0.033
γ_3	0.200	0.214	0.196	0.202			
ω_1	-0.200	-0.259	-0.231	-0.223			
ω_2	0.500	0.720	0.599	0.544	1.782	0.426	0.212
ω_3	1.000	1.400	1.162	1.102			

Note que para $n = 50$ as médias das estimativas dos coeficientes para modelagem da locação e escala já estão muito próximos aos valores verdadeiros, enquanto que as médias das estimativas dos coeficientes para modelagem da assimetria se aproximam dos valores verdadeiros somente para amostras maiores que $n = 50$. Importante ressaltar que os valores do EQM decrescem conforme o tamanho da amostra aumenta, como esperado. O código utilizado para essa simulação é similar ao da Seção 3.2.1, e está disponível no Apêndice C.

4 APLICAÇÃO EM DADOS REAIS

Neste trabalho, apresentamos aplicações da teoria discutida em diversos campos do conhecimento. A primeira, com os dados da Martin Marietta (Butler et al., 1990), pertencente à área da Economia. Já a aplicação com os dados de circunferência abdominal (Fellow et al., 2005) vem da área da Biologia, enquanto os de produção de leite são da Zoologia. Os códigos utilizados em todas as três aplicações são muito semelhantes, e podem ser resumidos sem muita perda de informação no *snippet* descrito no Apêndice D. Observe que a quantidade máxima de iterações do ciclo externo do algoritmo RS foi aumentada para 500. Isto foi feito pois, em alguns casos, o número padrão de iterações (20) não foi suficiente para convergência do algoritmo. Dessa forma, como o tempo de processamento foi baixo, conseguimos aumentar o número de iterações sem custo computacional. Além disso, é importante lembrar que o código fornecido constrói apenas os modelos explicados por todas as variáveis explicativas possíveis. Após cada um desses modelos listados, devemos verificar a significância dos coeficientes através da função *summary* e seguir manualmente o método descrito na Seção 2.6.

4.1 MARTIN MARIETTA

Martin Marietta é um conjunto de dados apresentado em Butler et al. (1990) para estudar a relação entre os retornos mensais da ação da empresa Martin Marietta e os retornos do índice CRSP da bolsa de valores de Nova Iorque. A variável resposta diz respeito à taxa de retorno mensal (incluindo os dividendos) da empresa Martin Marietta. Isso refere-se ao ganho total que um investidor recebe de um ativo ao longo de um mês, sendo ele derivado tanto da valorização do ativo quanto dos dividendos pagos nesse período. Já a variável explicativa representa os retornos do portfólio de mercado, isto é, é um indicador de desempenho da maioria das companhias listadas na bolsa de valores de Nova Iorque. O índice é operado pelo Center for Research in Securities Prices (CRSP), uma instituição associada à Booth School of Business da Universidade de Chicago. O valor do índice é calculado não somente pelo produto entre o valor das ações e a quantidade de ações no mercado, mas levando em consideração também os dividendos pagos pelas empresas. Ambas as taxas de retorno utilizadas são as excedentes, isto é, referem-se ao retorno do ativo além da taxa livre de risco. A proxy utilizada para a taxa livre de risco foi o rendimento de um título do Tesouro dos Estados Unidos em trinta dias (30-Day Treasury Bill rate).

Este conjunto de dados foi analisado por Butler et al. (1990) para ajustar uma regressão linear simples com erros gaussianos, enquanto Azzalini e Capitanio (2003) propuseram um modelo de regressão linear com erros assumindo uma ST. Além disso, Taylor e Verbyla (2004) propuseram uma modelagem conjunta dos parâmetros de locação

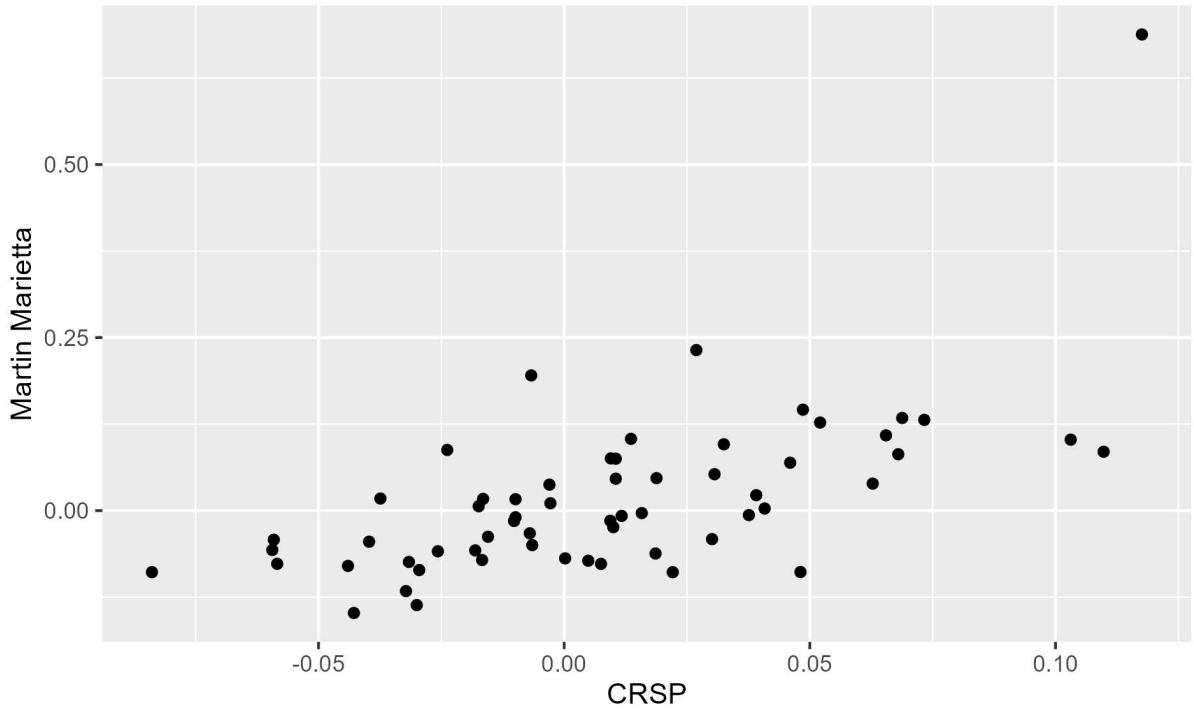


Figura 7: Gráfico de dispersão entre os retornos de Martin Marietta e os retornos do índice CRSP.

e escala da distribuição TF e Doğru e Arslan (2019) fizeram um comparativo da skew-t-normal e skew-laplace-normal modelando conjuntamente os parâmetros de locação, escala e assimetria considerando algoritmo EM para obter os estimadores de máxima verossimilhança dos parâmetros desses modelos.

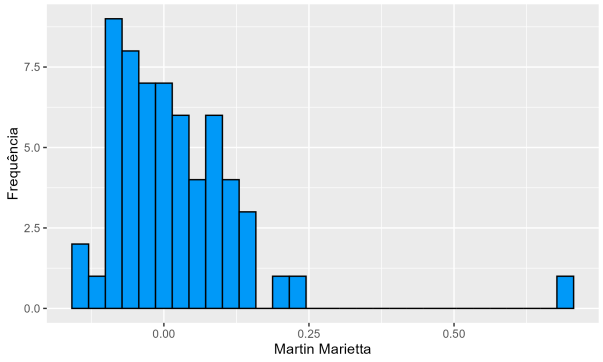
Conforme descrito em Pokharel et al. (2024), os rendimentos do mercado de ações muitas vezes apresentam assimetria, além de curtose e peso nas caudas maiores do que os de uma distribuição normal. Esses são alguns dos chamados fatos estilizados dos retornos financeiros encontrados na literatura empírica; veja Cont (2001) para uma descrição completa desses fatos.

Seguindo os resultados da literatura empírica, ajustamos o GAMLSS baseados nas duas distribuições apresentadas na Seção 2.2 (skew normal e skew t) e suas versões simétricas (NO e TF). Cabe mencionar que, até onde sabemos, nenhum trabalho modelou o parâmetro da curtose na skew-t usando variáveis explicativas para o conjunto de dados de Martin Marietta, conforme consideramos neste trabalho.

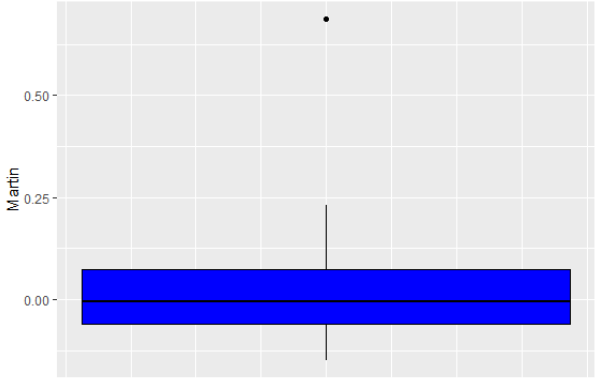
4.1.1 Análise exploratória dos dados

A Figura 7 apresenta o gráfico de dispersão entre os retornos de Martin Marietta e os retornos do índice CRSP, a Figura 8(a) apresenta o histograma dos retornos de Martin Marietta e a Figura 8(b) apresenta seu boxplot.

Conforme podemos ver na Figura 7, os retornos excedentes da empresa Martin



(a) Histograma da taxa de retorno da Martin Marietta.



(b) Boxplot da taxa de retorno da Martin Marietta.

Figura 8: Gráficos da taxa de retorno da Martin Marietta.

Marietta parecem seguir uma relação linear crescente com os do CRSP. Dessa forma, espera-se que, para todas as distribuições, o coeficiente da variável CRSP seja positivo na equação da locação. Além disso, por termos muitos valores negativos, esperamos que o intercepto da equação da locação seja negativo. Podemos notar também que os pontos parecem se dispersar mais conforme aumentam os valores de CRSP. Dessa forma, supõe-se que o valor do coeficiente de CRSP na equação da variância seja positivo. Além disso, como mostra o histograma da Figura 8(a), presume-se que as distribuições assimétricas sejam mais adequadas. Tais resultados estão de acordo com os encontrados na literatura estatística sobre a distribuição dos retornos de ativos financeiro. Por fim, devido à presença de um outlier (facilmente visualizado na Figura 8(b)), esperamos que as distribuições que tenham caudas mais pesadas se ajustem melhor a este conjunto de dados.

4.1.2 Escolha do modelo

Considerando cada uma das quatro distribuições da Seção 2 (NO, SN, TF, ST), estimamos todos os parâmetros do modelo GAMLSS em função de CRSP, e o chamamos de modelo completo. Em seguida, retiramos CRSP das equações dos parâmetros de acordo com a sua significância, e o chamamos de modelo final.

Uma vez que temos ajustado um modelo para cada uma das quatro distribuições, utilizamos os critérios AIC e BIC para selecionar o melhor modelo em termos de ajuste aos dados. A Tabela 8 apresenta os resultados de AIC e BIC para cada modelo final da distribuição considerada. Observe que para qualquer um dos critérios a distribuição SN apresentou um melhor ajuste, o que corrobora com algumas das análises exploratórias feitas na Seção 4.1.1, como a forte assimetria apresentada no histograma dos retornos de Martin Marietta (Figura 8(a)).

Modelo	p	$\ell(\hat{\theta})$	AIC	BIC
NO	4	-142.049	-134.0492	-125.6718
SN	5	-151.835	**-141.8347**	**-131.3629**
TF	4	-143.626	-135.6258	-127.2485
ST	6	-147.259	-135.2594	-122.6933

Tabela 8 – AIC e BIC dos modelos ajustados aos dados de Martin Marietta

O modelo final escolhido está descrito em (4.1). Note que o intercepto da equação da locação é negativo, e que o valor do coeficiente de CRSP na equação da escala foi positivo, resultados já esperados da análise exploratória.

$$\begin{cases} \text{MM} \sim SN(\mu_i, \sigma_i^2, \lambda_i), \\ \hat{\mu}_i = -0.05617, \\ \log(\hat{\sigma}_i^2) = -2.46023 + 12.94994\ CRSP_i, \\ \hat{\lambda}_i = 1.7730 + 54.7783\ CRSP_i \ \text{ e} \\ i = 1, \ldots, 60. \end{cases} \tag{4.1}$$

Os erros padrões associados às estimativas dos parâmetros do modelo (4.1) podem ser encontrados no Anexo A.

4.1.3 Análise dos resíduos

Como o parâmetro de locação não foi modelado por covariáveis, não usamos o gráfico de resíduos exatamente como descrito na Seção 2.8, já que o primeiro gráfico seria irrelevante. Os demais estão apresentados na Figura 9. Com base na Figura 9(a) vemos que a suposição de independência dos erros foi atendida, pois os resíduos parecem se comportar de maneira aleatória sem apresentar nenhuma tendência ao longo do tempo. Nas Figuras 9(b) e 9(c), vemos que a normalidade dos resíduos quantílicos não foi violada. É importante ressaltar que a reta ideal (em vermelho) se distancia dos pontos em alguns momentos, especialmente nas extremidades. Contudo, devemos ter em mente que a amostra é de tamanho 60, e, portanto, até mesmo para simulações esse resultado foi observado. Dessa forma, não descartamos a normalidade dos resíduos quantílicos, corroborando a qualidade do ajuste. Isso pode ser melhor visto nos worm plots da Figura 10, em que os resíduos estão bem distribuídos horizontalmente.

4.1.4 Interpretação dos parâmetros

Depois de verificada a qualidade do modelo final escolhido, o próximo passo então é interpretar os coeficientes do mesmo. Nesta seção, para simplificar a notação, desconsideramos o índice i.

Conforme o modelo selecionado, temos que o log do parâmetro de escala pode ser escrito como

$$log(\hat{\sigma}^2) = -2.46 + 12.95CRSP.$$

E, se CRSP aumentar em 0.2 unidades, temos que

$$log(\hat{\sigma}^{2*}) = -2.46 + (CRSP + 0.2)12.95 = 0.13 + 12.95CRSP.$$

E consequentemente, temos que

$$log(\hat{\sigma}^{2*}) - log(\hat{\sigma}^2) = 0.13 + CRSP12.95 - (-2.46 + CRSP12.95),$$

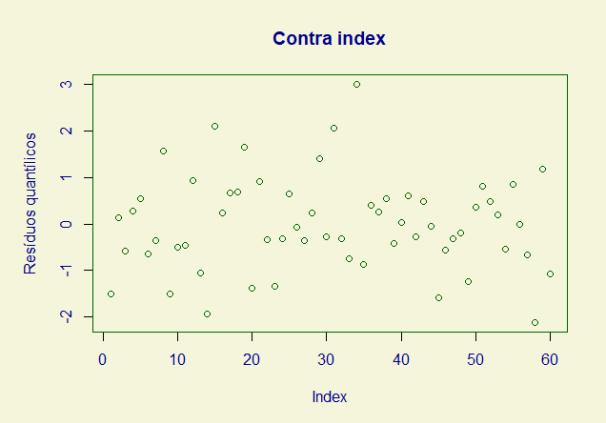
tal que,

$$log(\frac{\hat{\sigma}^{2*}}{\hat{\sigma}^2}) = 2.59.$$

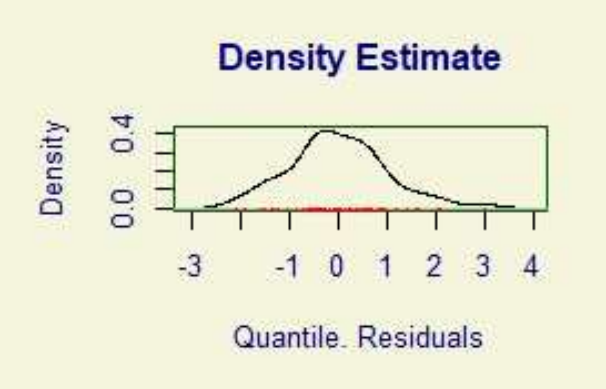
Concluindo, podemos escrever como

$$\frac{\hat{\sigma}^{2*}}{\hat{\sigma}^2} = exp(2.59) = 13.33,$$

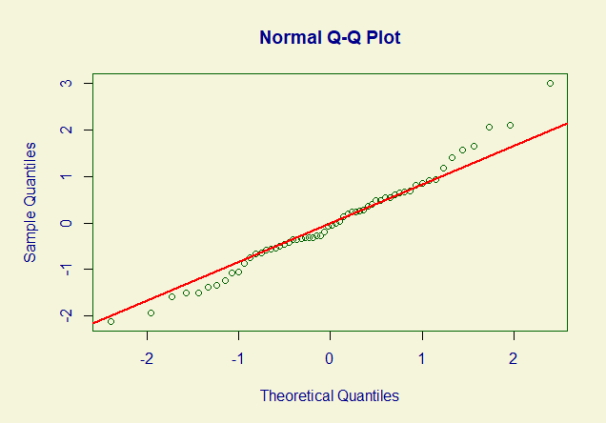
ou seja, o aumento de 0.2 unidades no CRSP (o que é aproximadamente a maior variação apresentada no conjunto de dados), causa um aumento de 13.33 vezes no parâmetro de escala estimado. Consequentemente, esse aumento impacta na escala estimada em $\sqrt{13.33} = 3.65$ vezes.



(a) Resíduos quantílicos ao longo de tempo.



(b) Densidade estimada.

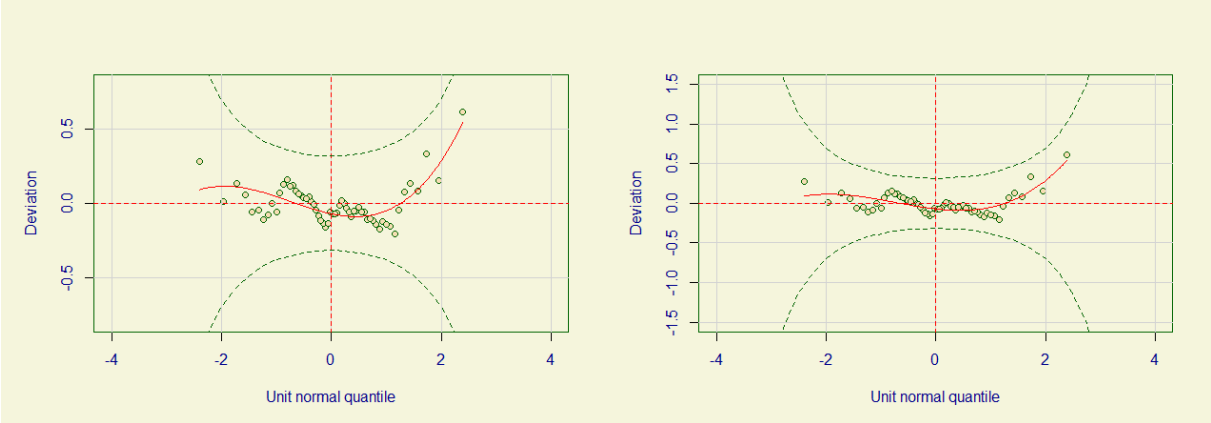


(c) Q-Q Plot.

Figura 9: Gráficos dos resíduos quantílicos.

Analogamente, podemos avaliar o aumento de 0.2 unidade no CRSP no parâmetro de assimetria estimado. Assim, segue que

$$\hat{\lambda} = 1.77 + CRSP54.78 \text{ e}$$
$$\hat{\lambda}^* = 1.77 + (CRSP + 0.2)54.78.$$



(a) Limites do gráfico mantidos como padrão. (b) Limites do gráfico ampliados.

Figura 10: Worm plot dos resíduos quantílicos.

Então,

$$\hat{\lambda}^* - \hat{\lambda} = (0.2)(54.74) = 10.95,$$

ou seja, o aumento de 0.2 unidades no CRSP causa um aumento de 10.95 unidades no λ estimado.

Como mostrado na Seção 2.2.1, tanto σ quanto λ impactam no valor da média e da variância da distribuição. Assim, é difícil interpretar essas mudanças numericamente, fazendo necessário o uso de gráficos para compreendermos melhor a diferença das distribuições estimadas ao longo dos valores de CRSP. A Figura 11 mostra em vermelho o valor da média estimada para cada ponto, e em preto a curva da função densidade de probabilidade da SN. Veja que quanto maior o valor de CRSP, mais dispersas são as distribuições, e mais assimétricas à direita também. Além disso, há um claro aumento na média das distribuições, indicando então que altos valores da taxa de retorno do CRSP estão correlacionados com maiores taxas de retorno para Martin Marietta, com o detalhe de que a variabilidade também aumenta, tendo uma vasta gama de possibilidades de valores acima da média.

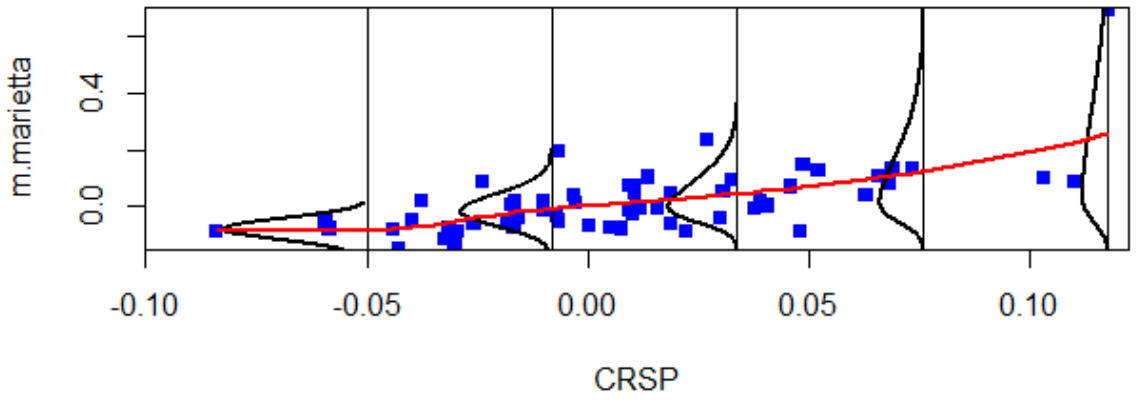


Figura 11: Gráfico das densidades estimadas nos quantis 0, 0.25, 0.50, 0.75 e 1 com uma curva indicando a média da distribuição nestes pontos.

4.2 CIRCUNFERÊNCIA ABDOMINAL

O conjunto de dados aqui utilizado foi o apresentado por Fellow et al. (2005). Nele encontramos duas variáveis, uma referente ao tamanho da circunferência abdominal, em milímetros, de 610 fetos, e a outra, que será usada como explicativa, diz respeito a idade gestacional do feto em semanas. A circunferência varia de 56 a 404 milímetros, enquanto a idade varia de 12.29 a 42.43 semanas.

No mesmo contexto, mas com outro conjunto de dados, Wright e Royston (2008) encontrou caudas pesadas na distribuição da circunferência abdominal, indicando que assumir normalidade dos dados traria viés na estimação dos quantis extremos, e, portanto, deve-se atentar mais aos maiores momentos da distribuição.

4.2.1 Análise exploratória

A Figura 12 mostra como a circunferência abdominal se comporta em função da idade gestacional, enquanto a Figura 13 mostra o histograma dos valores da circunferência abdominal. Com base no gráfico de dispersão da Figura 12, é razoável supor que o parâmetro de locação será explicado com base na idade gestacional, já que a circunferência abdominal claramente apresenta uma relação linear com ela. Além disso, o gráfico nos indica um aumento na dispersão dos pontos conforme aumenta a idade gestacional. Assim, esperamos que essa variável explicativa também seja mantida na modelagem do parâmetro de escala.

Fundamentado no histograma da Figura 13 (coeficiente de Pearson amostral da curtose igual a 1.91) e nos comentários feitos por Wright e Royston (2008), podemos supor que distribuições assimétricas e com caudas mais leves devem se ajustar melhor aos dados.

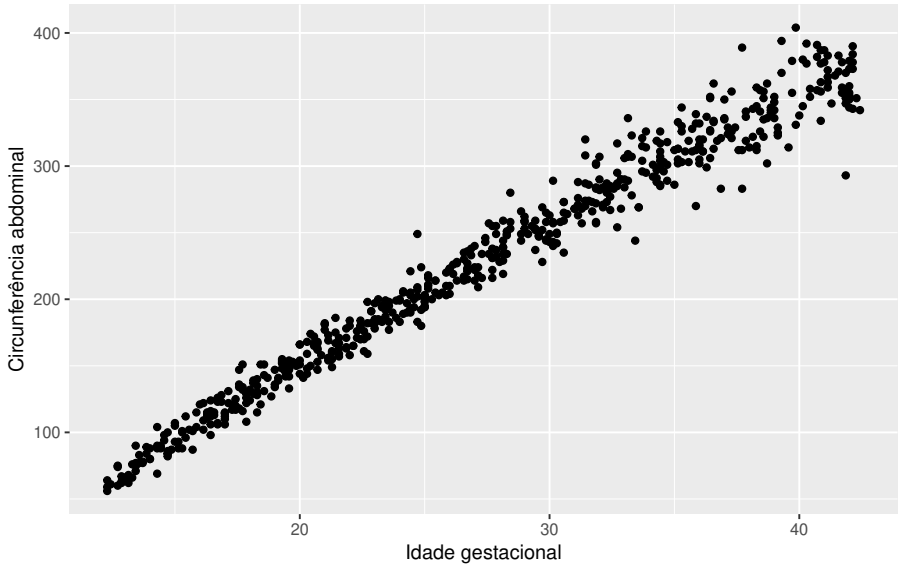


Figura 12: Gráfico de dispersão dos dados de circunferência abdominal.

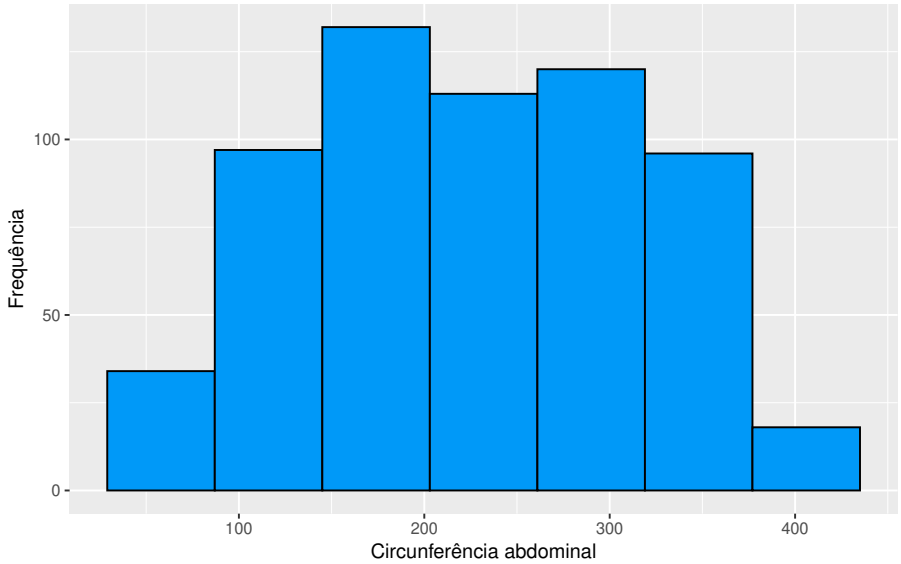


Figura 13: Histograma da circunferência abdominal.

Assim, dentre as quatro distribuições testadas, esperamos que a skew-normal seja a melhor em termos de ajuste aos dados.

4.2.2 Escolha do modelo

Como feito na seção anterior, considerando as distribuições da Seção 2.2 e suas versões simétricas (NO, SN, TF, ST), estimamos todos os parâmetros do GAMLSS em função da variável explicativa idade gestacional do feto em semanas, e o chamamos de modelo completo. Em seguida, a seleção da variável foi feita, considerando o nível de significância de 5%, para cada umas das equações dos parâmetros e obtemos o modelo final sob cada distribuição estudada. Os resultados do AIC e BIC de cada um dos modelos finais é apresentado na Tabela 9 a seguir.

Modelo	p	$\ell(\hat{\theta})$	AIC	BIC
NO	4	4853.18	4861.184	4878.838
SN	6	4796.77	4808.772	**4835.253**
TF	5	4847.37	4857.375	4879.442
ST	7	4794.34	**4808.341**	4839.236

Tabela 9 – AIC e BIC dos modelos ajustados

Temos dois possíveis modelos para serem escolhidos, um com distribuição skew-normal, e outro com distribuição skew-t. Observe que o AIC indica que o modelo ST se ajusta melhor aos dados enquanto que o BIC aponta para o modelo SN. Ambos estão descritos nas equações (4.2) e (4.3).

$$\begin{cases} \text{Circunferência} \sim SN(\mu_i, \sigma_i^2, \lambda_i), \\ \hat{\mu}_i = -78.36 + idade_i 10.95, \\ \log(\hat{\sigma}_i^2) = 2.18 + idade_i 0.02 \text{ e} \\ \hat{\lambda}_i = 5.28 + idade_i(-0.14). \end{cases} \tag{4.2}$$

$$\begin{cases} \text{Circunferência} \sim ST(\mu_i, \sigma_i^2, \lambda_i, \tau_i), \\ \hat{\mu}_i = -78.10 + idade_i 10.96, \\ \log(\hat{\sigma}_i^2) = 2.12 + idade_i 0.02, \\ \hat{\lambda}_i = 5.06 - idade_i 0.14 \text{ e} \\ \log(\hat{\tau}_i) = 3.00. \end{cases} \tag{4.3}$$

Para decidir qual modelo será o escolhido, iremos analisar os resíduos quantílicos. Assim, comparando as Figuras 14 e 15, podemos notar que nenhuma apresenta sinais de heterocedasticidade, mas que os valores dos resíduos quantílicos da SN parecem ser mais assimétricos à esquerda, evidenciado pela fato da escala do gráfico ir até -4. Ademais, verificamos que o Q-Q Plot do modelo ST é mais alinhado, principalmente nas extremidades, à reta ideal.

Por fim, comparando os worm plots das Figuras 16 e 17, podemos notar que ambos não apresentam problemas, apesar do da ST apresentar pontos mais próximos da reta horizontal. Assim, a análise dos resíduos não tornou muito clara a escolha da distribuição a ser utilizada, e, portanto, utilizaremos outros critérios. Através da análise exploratória dos dados, vimos que a circunferência abdominal segue distribuição com caudas leves, indicando que a SN é mais indicada. Além disso, conforme comentado na Seção 2.7, o BIC é um critério mais rigoroso, tornando o seu modelo mais parcimonioso. Dessa forma, opta-se por utilizar o modelo (4.2). Os erros padrões associados às estimativas dos parâmetros do modelo (4.2) podem ser encontrados no Anexo A.

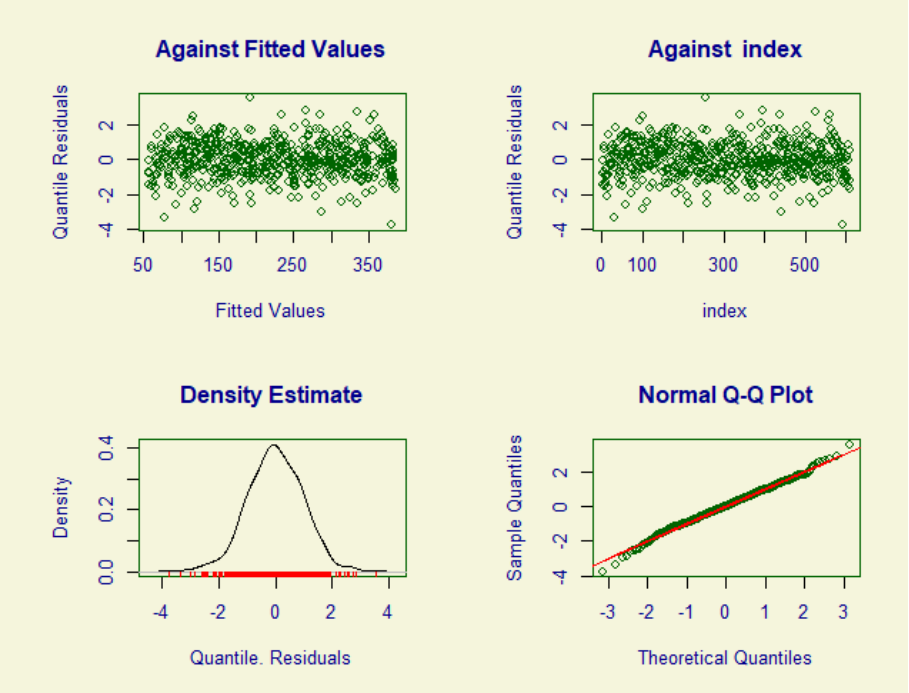


Figura 14: Resíduos do modelo skew-normal.

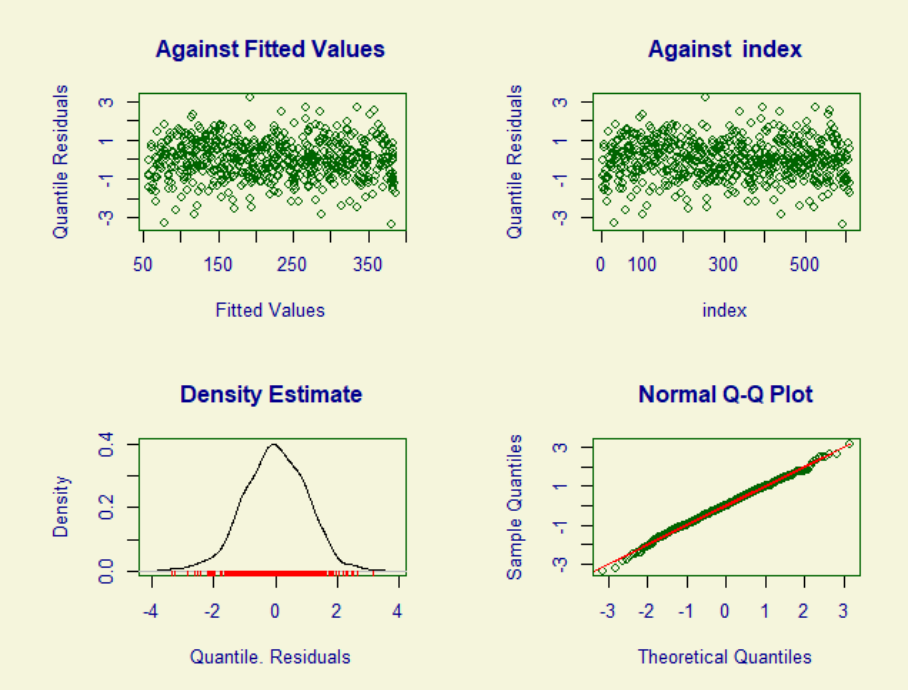
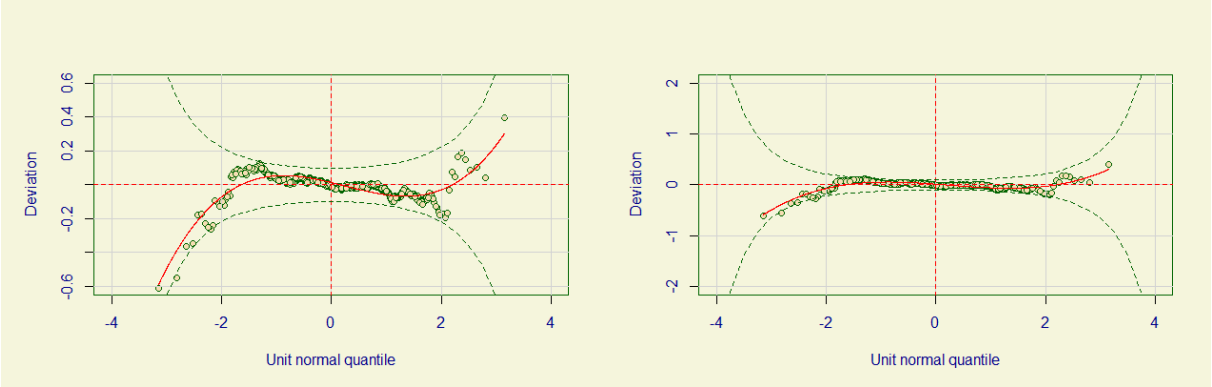


Figura 15: Resíduos do modelo skew-t.

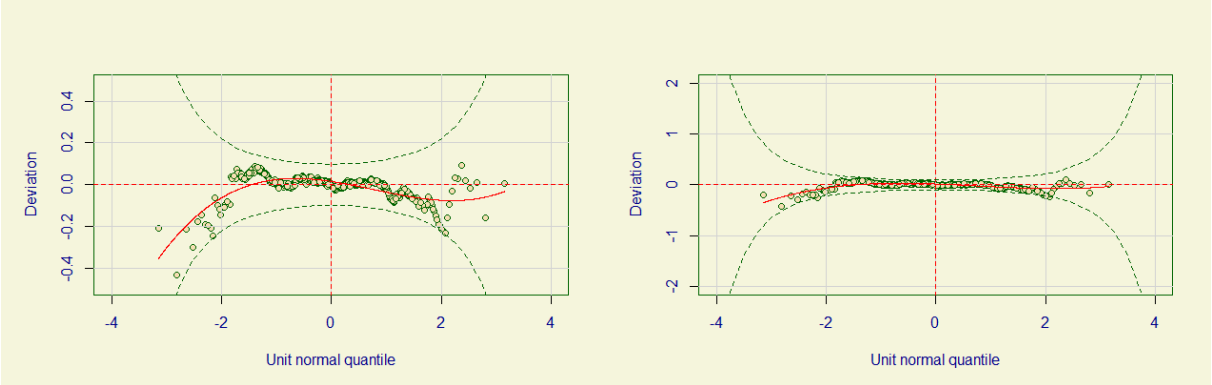
4.2.3 Interpretação dos parâmetros

Nesta seção, para simplificar a notação, desconsideramos o índice i. Conforme o modelo selecionado, podemos avaliar o impacto do aumento de 1 semana na idade no parâmetro de locação, para tanto, segue que



(a) Limites do gráfico mantidos como padrão. (b) Limites do gráfico ampliados.

Figura 16: Worm plot do modelo skew-normal.



(a) Limites do gráfico mantidos como padrão. (b) Limites do gráfico ampliados.

Figura 17: Worm plot do modelo skew-t.

$$\hat{\mu} = -78.36 + idade 10.95 \text{ e}$$
$$\hat{\mu}^* = -78.36 + (idade + 1)10.95.$$

Então,

$$\hat{\mu}^* - \hat{\mu} = 10.95.$$

Logo, a estimativa da locação aumenta em 10.95 milímetros se a idade gestacional durar uma semana a mais.

Similarmente, temos que

$$log(\hat{\sigma}^2) = 2.18 + idade 0.02,$$

e, se a idade gestacional aumentar em uma semana, obtemos a estimativa do parâmetro de escala cujo $\log(\hat{\sigma}^2)$ é dado por

$$log(\hat{\sigma}^{2*}) = 2.18 + (idade + 1)0.02 = 2.2 + idade 0.02.$$

E consequentemente, temos que

$$log(\hat{\sigma}^{2*}) - log(\hat{\sigma}^2) = 2.2 + idade0.02 - (2.18 + idade0.02) \text{ e}$$

$$log(\frac{\hat{\sigma}^{2*}}{\hat{\sigma}^2}) = 0.02,$$

que podemos reescrever como

$$\frac{\hat{\sigma}^{2*}}{\hat{\sigma}^2} = exp(0.02) = 1.02.$$

Concluindo, o aumento de uma semana na idade gestacional do feto (lembrando que a maior variação vista nos dados é de 30.14 semanas), causa um aumento de 1.02 vezes na estimativa de σ^2. Consequentemente, esse aumento impacta na estimativa σ em $\sqrt{1.02} = 1.01$ vezes.

Analogamente, podemos avaliar o aumento de uma semana de idade gestacional na estimativa da assimetria. Assim, segue que

$$\begin{cases} \hat{\lambda} = 5.28 - idade0.14 \\ \hat{\lambda}^* = 5.28 - (idade + 1)0.14. \end{cases} \tag{4.4}$$

Então,

$$\hat{\lambda}^* - \hat{\lambda} = -0.14,$$

ou seja, o aumento de uma semana na idade gestacional causa uma diminuição de 0.14 na estimativa de λ, tornando a distribuição simétrica por volta da 36º semana de gestação, e negativa posteriomente.

Como feito nos dados do Martin Marietta, é necessário o uso de gráficos para compreendermos melhor a diferença das distribuições estimadas ao longo dos valores da idade gestacional. A Figura 18 mostra claramente que a média da circunferência abdominal aumenta com a idade gestacional, mas não necessariamente de forma linear, já que nas idades maiores há uma tendência de estabilização da circuferência abdominal. Além disso, conforme descrito matematicamente anteriormente, a assimetria das distribuições ajustadas vai sempre diminuindo, chegando a se tornar uma distribuição muito simétrica por volta da idade de 35 semanas. Por fim, podemos notar que com maiores idades temos maiores dispersões nas distribuições.

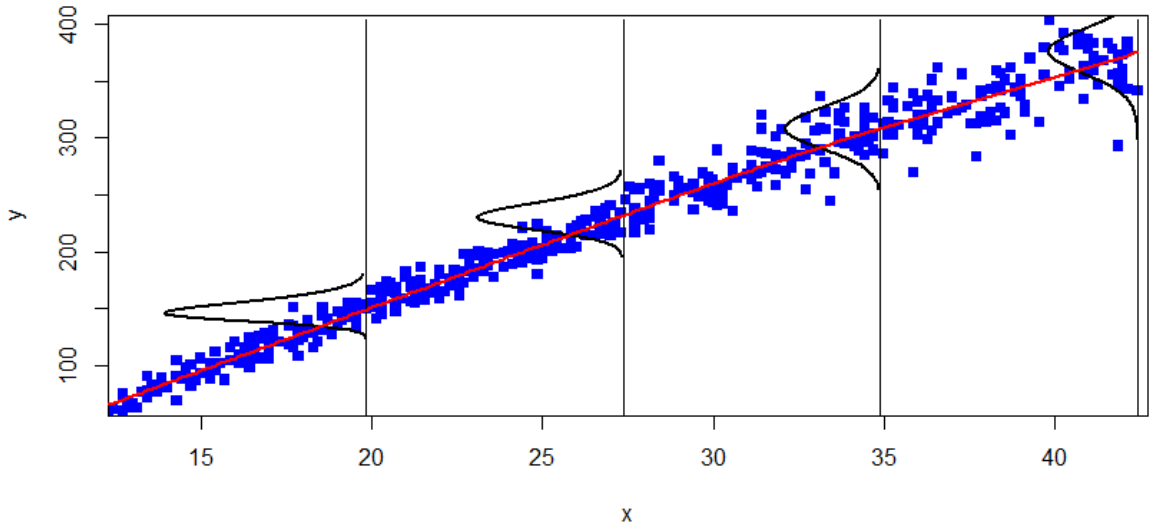


Figura 18: Gráfico das densidades estimadas nos quantis 0, 0.25, 0.50, 0.75 e 1 com uma curva indicando a média da distribuição nestes pontos.

4.3 PRODUÇÃO DE LEITE

Este conjunto de dados possui 4 variáveis e é utilizado pela Embrapa em parceria com a Associação Brasileira de Criadores de Bovinos da Raça Holandesa (ABCBRH) para realização das avaliações genéticas oficiais da raça Holandesa no Brasil, as quais estão disponíveis em Gado Holandês (2024). A variável resposta diz respeito à quantidade de leite total produzido, em quilogramas, por vacas da raça Holandesa durante uma lactação. As demais, explicativas, referem-se à lactação em que a produção foi alcançada, ao ano em que o parto ocorreu e à idade, em categoria, da vaca à época do parto. Lactação com valor 1 significa que estamos nos referindo ao primeiro período (que dura em média 305 dias) em que a vaca produziu leite até secar. Lactação 2 significa que a vaca pariu novamente e voltou a produzir leite por mais um ciclo. Essa mesma lógica se aplica à 3º lactação, valor máximo desse conjunto de dados. Idade com valor um significa que a vaca tinha até dois anos de idade. Para cada categoria a mais são adicionados 3 meses. Assim, idade igual a dois significa que a vaca tinha entre dois anos e dois anos e três meses. Idade com valor três significa que a vaca tinha entre dois anos e três meses e dois anos e seis meses, e assim por diante. Com o intuito de analisar este conjunto de dados no contexto do GAMLSS paramétrico, quando um animal tinha uma produção total calculada em mais de uma lactação (medidas repetidas), o procedimento adotado foi sortear para manter apenas uma observação (produção) por vaca. Dessa forma, o conjunto de dados é composto por 6099 observações.

Usualmente, ao longo da lactação, a quantidade de peso é medida apenas algumas vezes. Dessa forma, divide-se os 305 dias esperados da lactação do animal em dez intervalos.

Nem todos os animais passam por todos os controles leiteiros, mas aqui foram selecionados apenas as vacas que tiveram o leite pesado dez vezes durante a lactação.

Para calcular a quantidade total de leite produzida vários modelos já foram utilizados, para mais detalhes, veja Guimarães et al. (2006). Contudo, neste trabalho, a fim de simplificar o cálculo da quantidade total de leite, foram utilizadas médias. Dessa forma, se na primeira pesagem a vaca produziu 20kg, assumimos que, durante todo esse período de dez dias (isto é, do dia 6 ao 15, já que os primeiros cinco dias de produção são ignorados), ela produziu 200kg. Já se na segunda pesagem foram medidos 30kg, então consideramos que durante os dias 15 e 25 (que abrangem o segundo controle leiteiro), a vaca produziu em média $\frac{20+30}{2} = 25Kg$ por dia, isto é, 250kg durante o segundo controle leiteiro. Ao final, foi considerado que a última produção registrada da vaca foi mantida por mais 15 dias, isto é, foi concedido um crédito de quinze dias ao animal por não sabermos se ela de fato secou após o registro ou se continou produzindo.

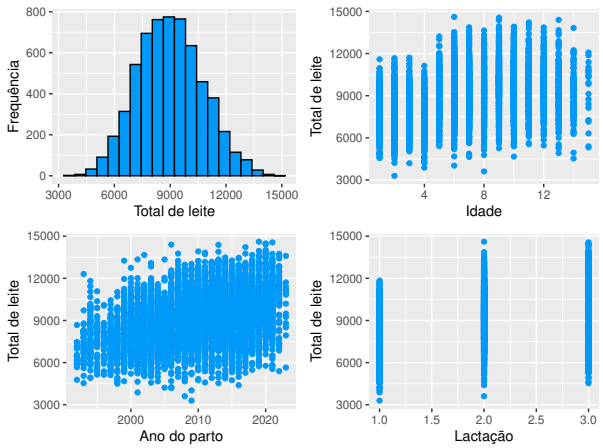


Figura 19: Características dos dados.

4.3.1 Análise exploratória

Com o histograma de frequência disponível na Figura 19, podemos notar que a distribuição normal provavelmente se ajustará bem aos dados, apesar de, talvez tenha uma leve assimetria à direita (assimetria amostral de 0.14), o que nos leva a pensar que a skew-normal pode ser uma distribuição adequada aos dados. Além disso, analisando os demais gráficos, vemos que a relação entre leite total produzido e as covariáveis é crescente, então esperamos que os coeficientes dessas variáveis na equação da locação sejam positivos. Adicionalmente, parece que o aumento na lactação e ano do parto fazem a variável resposta variar mais, levando à hipótese de que elas poderão ter coeficiente positivo na equação da escala.

4.3.2 Escolha das variáveis e modelo

Seguindo o mesmo procedimento adotado nas aplicações anteriores, para cada distribuição (NO, SN, TF,e ST), ajustamos o modelo completo, isto é, incluindo todas as variáveis explicativas nas equações dos parâmetros. Em seguida, em cada equação do parâmetro, considerando as etapas descritas na Seção 2.6, retiramos as variáveis explicativas cujos coeficientes não se mostraram significantes ao nível de 5%. A Tabela 10 apresenta o AIC e BIC dos modelos ajustados para cada distribuição depois de excluídas as variáveis explicativas não significantes em cada equação dos parâmetros. De acordo com os critérios AIC e BIC, os modelos que melhor se ajustam a este conjunto de dados são o GAMLSS sob distribuição ST e NO, descrito em (4.5) e (4.6) respectivamente. Dessa forma, a análise de resíduos se faz necessária para a escolha da distribuição.

$$\begin{cases} \text{Leite} \sim \text{ST}(\mu_i, \sigma_i^2, \lambda_i, \tau_i), \\ \mu_i = -2.43\ 10^5 + 83 idade + 405.1\ \text{I}(lactacao_i = 2) + 646.2\ \text{I}(lactacao_i = 3) + 124.7\text{anoParto}_i, \\ \log(\sigma_i^2) = 7.144 + 0.283\ \text{I}(lactacao_i = 2) + 0.290\ \text{I}(lactacao_i = 3), \\ \lambda_i = 116.31 + -0.095 idade + 1.03\ \text{I}(lactacao_i = 2) + 1.15\ \text{I}(lactacao_i = 3) - 0.058\text{anoParto}_i, \\ log(\tau_i) = 322.36 - 1.405 idade + 9.59\ \text{I}(lactacao_i = 2) + 11.32\ \text{I}(lactacao_i = 3) - 0.154\text{anoParto}_i, \\ i = 1, \ldots, 6099, \end{cases} \tag{4.5}$$

onde I(.) é a função indicadora, ou seja, I($lactacao_i = 2$) assume 1 caso o valor de lactação da i-ésima observação seja 2, e assume 0 caso contrário. O mesmo raciocínio se aplica a I($lactacao_i = 3$).

$$\begin{cases} \text{Leite} \sim \text{NO}(\mu_i, \sigma_i^2), \\ \mu_i = -1.474\ 10^5 + 1.386\ 10^3\ \text{I}(lactacao_i = 2) + 1.737\ 10^3\ \text{I}(lactacao_i = 3) + 77.3\text{anoParto}_i, \\ \log(\sigma_i^2) = -9.24 + 0.02\text{idade} + 0.103\ \text{I}(lactacao_i = 2) + 0.081\ \text{I}(lactacao_i = 3) + 0.008\text{anoParto}_i, \\ i = 1, \ldots, 6099. \end{cases} \tag{4.6}$$

Modelo	p	$\ell(\hat{\theta})$	AIC	BIC
NO	9	105886	105903.6	**105964.0**
SN	15	105871	105901.2	106002.0
TF	14	105881	105909.3	106003.3
ST	18	105845	**105880.9**	106001.8

Tabela 10 – AIC e BIC dos modelos ajustados.

4.3.3 Analise dos resíduos

Conforme vemos nas Figuras 20 e 21, os resíduos não parecem ter um padrão, e a normalidade dos resíduos quantílicos parece ser algo muito plausível. Por outro lado, pelo worm plot da Figura 22, podemos ver que os resíduos quantílicos do modelo normal parecem seguir a reta horizontal esperada, apesar de alguns pontos ultrapassarem a banda inferior de confiança no lado esquerdo do gráfico. Analogamente, o worm plot do modelo skew-t apresentado na Figura 23 mostra resíduos que ultrapassam a banda de confiança nas pontas e no centro do grafico, indicando um ajuste pior. Essa análise, aliada ao fato do modelo NO ser mais parcimonioso, justifica a escolha desse modelo. Os erros padrões associados às estimativas dos parâmetros do modelo (4.6) podem ser encontrados no Anexo A.

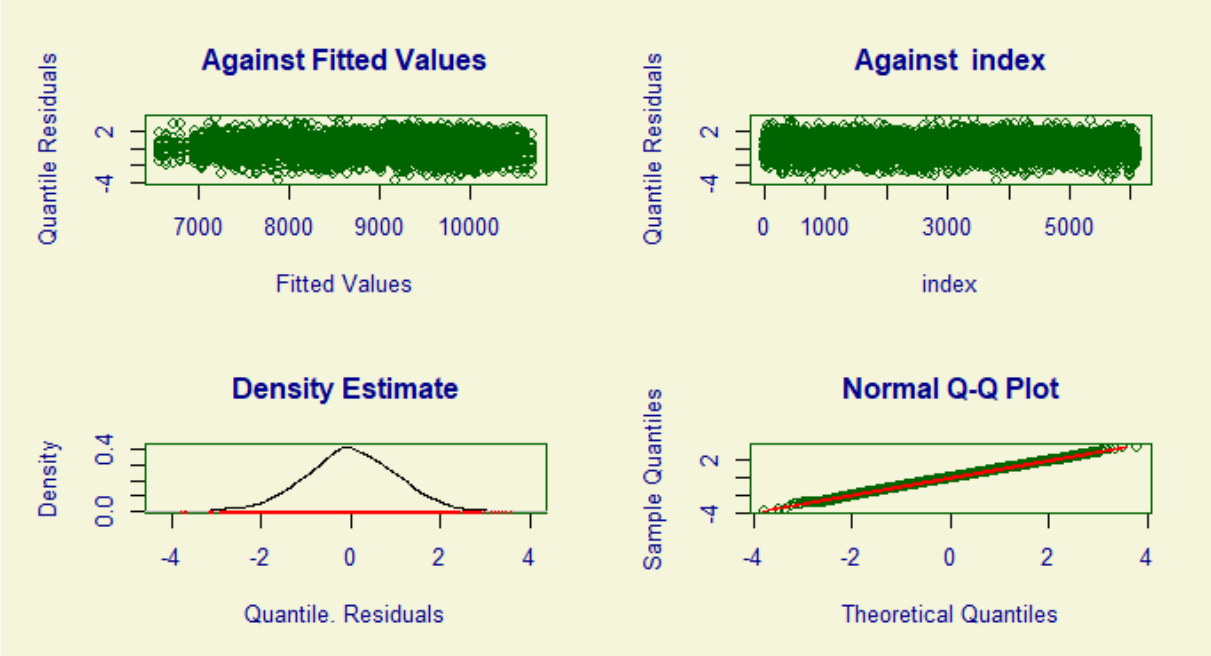


Figura 20: Gráficos de resíduos do modelo ajustado NO.

4.3.4 Interpretação dos parâmetros

Nesta seção, para simplificar a notação, desconsideramos o índice i. Analogamente ao feito nas aplicações anteriores, interpretamos o modelo GAMLSS sob distribuição normal ajustado. Dessa forma, entende-se vacas na segunda lactação tendem a produzir, em média, 138.6kg de leite a mais do que vacas na primeira lactação. De forma similar, podemos dizer que vacas na terceira lactação tendem a produzir, em média, 173.7kg de leite a mais do que vacas na primeira lactação, que também pode ser entendido como 35.1kg a mais que vacas na primeira lactação. Além disso, o aumento de um ano no ano do parto aumenta a produção em 77.3kg, indicando que vacas com parto mais recente produzem mais leite.

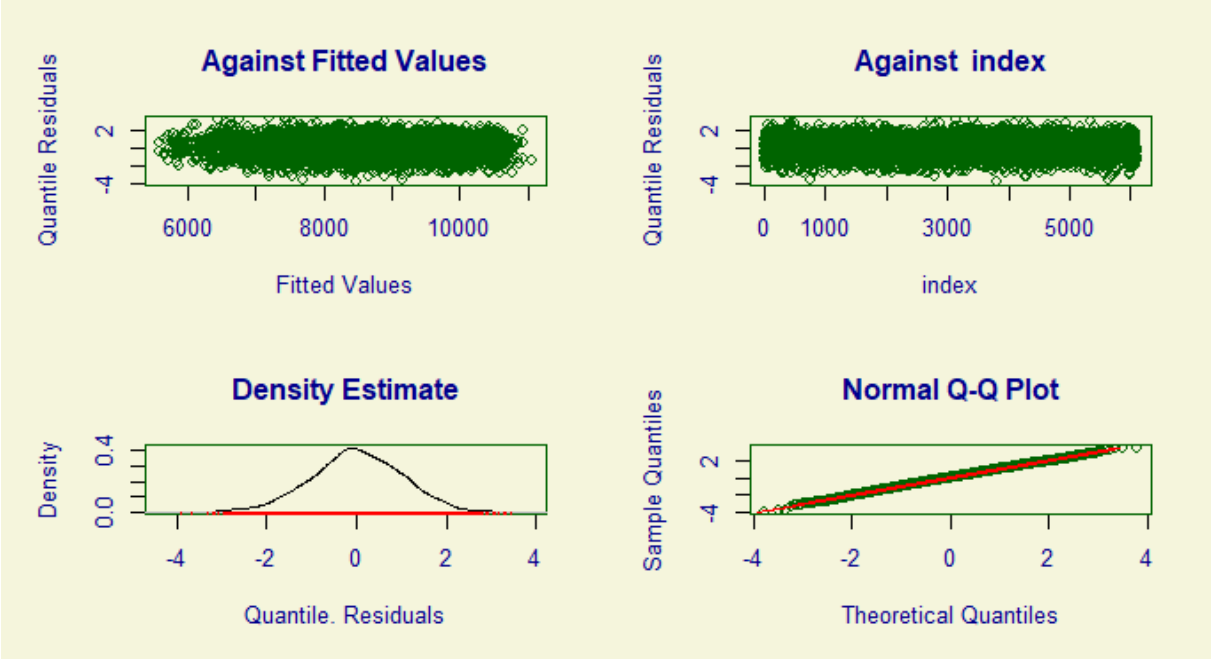
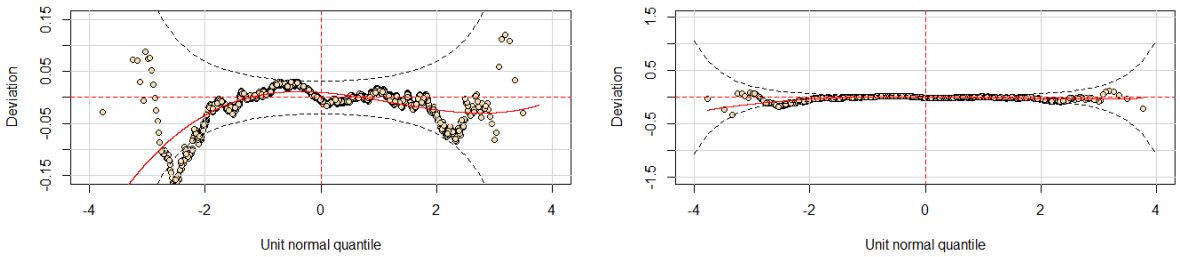


Figura 21: Gráficos de resíduos do modelo ajustado ST.

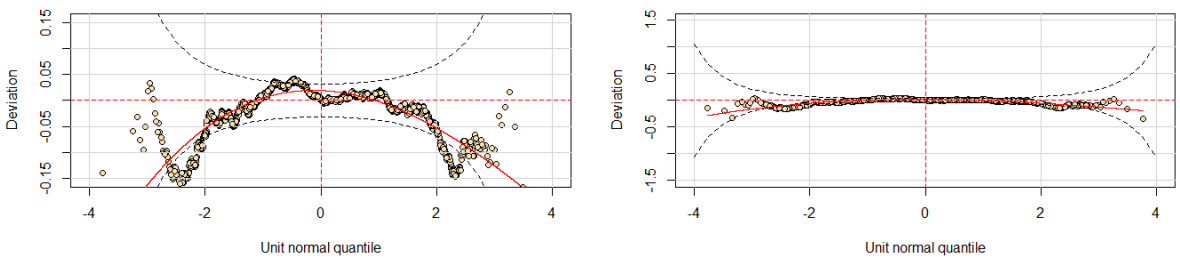


(a) Limites do gráfico mantidos como padrão. (b) Limites do gráfico ampliados.

Figura 22: Worm plot do modelo ajustado para NO.

Podemos interpretar também a mudança na variância estimada. Dessa forma, o aumento em um ano na idade da vaca aumenta a variância estimada em 1.021 vezes, enquanto que o aumento de uma unidade no ano do parto aumenta em 1.008 vezes. Analogamente, a lactação 2 possui aumento na variância estimada (com relação a lactação 1) de 1.109 vezes, e a lactação 3 aumento de 1.0839 vezes. Interessante notar que ao compararmos a terceira com a segunda lactação, observamos diminuição na variância estimada.

É importante ressaltar que este trabalho traz contribuições ao aplicar o modelo GAMLSS na análise de um conjunto de dados sobre produção de leite. Até onde foi possível verificar, essa abordagem é inédita nesse contexto. As contribuições realizadas não apenas ampliam o entendimento sobre a variabilidade e assimetria nos dados, mas também destacam a aplicabilidade do modelo nessa área.



(a) Limites do gráfico mantidos como padrão. (b) Limites do gráfico ampliados.

Figura 23: Worm plot do modelo ajustado para ST.

5 CONCLUSÕES

O presente trabalho estudou o modelo GAMLSS paramétrico e o aplicou utilizando o software R, especificamente o pacote gamlss, implementado por Rigby e Stasinopoulos (2005). O GAMLSS surge da necessidade de obter modelos de regressão para dados que seguem distribuições além da família exponencial. Tal metodologia ainda permite ajustar os parâmetros de locação, escala e forma em função de covariáveis.

Para mostrar a aplicabilidade da metodologia GAMLSS, análise de dados reais nos contextos de Medicina, Economia e Zootecnia foram realizadas. Quando o conjunto de dados apresentou certas especificidades como não normalidade, assimetria e/ou caudas pesadas, as distribuições assimétricas skew-normal e skew-t foram consideradas no contexto do GAMLSS.

Por fim, conclui-se que o presente trabalho atingiu o objetivo proposto de estudar o GAMLSS paramétrico, buscando entender a estimação das relações paramétricas entre a variável resposta e as explicativas, a escolha das distribuições, a seleção das variáveis explicativas e o diagnóstico do modelo.

5.1 TRABALHOS FUTUROS

Na perspectiva de trabalhos futuros, podemos considerar modelos semi-paramétricos ou não paramétricos, com as mais diversas funções de suavização nos cenários da análise de dados reais. Em especial, na aplicação em dados da zootecnia, consideramos como variável resposta a produção total de leite das vacas para garantir independência das observações. No entanto, como análise futura para este conjunto de dados, há a possibilidade de se trabalhar com as medidas repetidas, disponíveis na base de dados, considerando o GAMLSS com a inclusão de efeitos aleatórios, por exemplo.

REFERÊNCIAS

AZZALINI, A. A class of distributions which includes the normal ones. **Scandinavian Journal of Statistics**, v. 12, p. 171–178, 1985. Disponível em: <https://api.semanticscholar.org/CorpusID:116032535>.

AZZALINI, A.; CAPITANIO, A. Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew T Distribution. **Journal of the Royal Statistical Society Series B**, v. 65, p. 367–389, fev. 2003. DOI: 10.1111/1467-9868.00391.

BUTLER, R. J.; MCDONALD, J. B.; NELSON, R. D.; WHITE, S. B. Robust and Partially Adaptive Estimation of Regression Models. **The Review of Economics and Statistics**, The MIT Press, v. 72, n. 2, p. 321–327, 1990. ISSN 00346535, 15309142. Disponível em: <http://www.jstor.org/stable/2109722>. Acesso em: 11 jun. 2024.

BUUREN, S. van; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in Medicine**, v. 20, p. 1259–1277, 2001. DOI: 10.1002/sim.746. Disponível em: <https://api.semanticscholar.org/CorpusID:14571841>.

CONT, R. Empirical properties of asset returns: stylized facts and statistical issues. **Quantitative Finance**, v. 1, p. 223–236, 2001. Disponível em: <https://api.semanticscholar.org/CorpusID:5625400>.

DOĞRU, F. Z.; ARSLAN, O. Joint Modelling of the Location, Scale and Skewness Parameters of the Skew Laplace Normal Distribution. **Iranian Journal of Science and Technology, Transactions A: Science**, v. 43, p. 1249–1257, mar. 2019. DOI: 10.1007/s40995-018-0598-5.

DUNN, P. K.; SMYTH, G. K. Randomized Quantile Residuals. **Journal of Computational and Graphical Statistics**, ASA Website, v. 5, n. 3, p. 236–244, 1996. DOI: 10.1080/10618600.1996.10474708.

FELLOW, L.; LABORATORY, D.; MIDWIFE, A. Charts of fetal size: 3. Abdominal measurements. **BJOG: An International Journal of Obstetrics Gynaecology**, v. 101, p. 125–131, ago. 2005. DOI: 10.1111/j.1471-0528.1994.tb13077.x.

GADO HOLANDÊS, Associação Brasileira de. **Página inicial**. [S.l.: s.n.], 2024. https://www.gadoholandes.com.br/. Acesso em: 23 ago. 2024.

GALTON, F. Regression Towards Mediocrity in Hereditary Stature. **The Journal of the Anthropological Institute of Great Britain and Ireland**, [Royal Anthropological Institute of Great Britain e Ireland, Wiley], v. 15, p. 246–263, 1886. ISSN 09595295. Disponível em: <http://www.jstor.org/stable/2841583>. Acesso em: 23 jun. 2024.

GUIMARÃES, V.; RODRIGUES, M.; SARMENTO, J.; ROCHA, D. Utilização de funções matemáticas no estudo da curva de lactação em caprinos. **Revista Brasileira De Zootecnia-brazilian Journal of Animal Science - REV BRAS ZOOTECN**, v. 35, p. 535–543, abr. 2006. DOI: 10.1590/S1516-35982006000200028.

HASTIE, T.J.; TIBSHIRANI, R.J. **Generalized Additive Models**. [S.l.]: Taylor & Francis, 1990. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN 9780412343902. Disponível em: <https://books.google.com.br/books?id=qa29r1Ze1coC>.

INTERNATIONAL MONETARY FUND. **United States: Financial Sector Assessment Program-Stress Testing-Technical Notes**. [S.l.], 2015. (IMF Country Report, 2015/173). Disponível em: <https://www.imf.org/en/Publications/CR/Issues/2016/12/31/United-States-Financial-Sector-Assessment-Program-Stress-Testing-Technical-Notes-43058>.

LI, H.; WU, L. Joint modelling of location and scale parameters of the skew-normal distribution. **Applied Mathematics-A Journal of Chinese Universities**, v. 29, p. 265–272, set. 2014. DOI: 10.1007/s11766-014-2916-9.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General)**, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238. Disponível em: <http://www.jstor.org/stable/2344614>. Acesso em: 23 jun. 2024.

POKHAREL, J. K.; ARYAL, G.; KHANAL, N.; TSOKOS, C. P. Probability Distributions for Modeling Stock Market Returns—An Empirical Inquiry. **International Journal of Financial Studies**, MDPI, v. 12, n. 2, p. 1–27, 2024.

POSIT TEAM. **RStudio: Integrated Development Environment for R**. Boston, MA, 2023. Disponível em: <http://www.posit.co/>.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>.

RAMIRES, T.; NAKAMURA, L.; RIGHETTO, A.; PESCIM, R.; MAZUCHELI, J.; CORDEIRO, G. A new semiparametric Weibull cure rate model: fitting different behaviors within GAMLSS. **Journal of Applied Statistics**, v. 46, p. 1–17, mai. 2019. DOI: 10.1080/02664763.2019.1611748.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized Additive Models for Location, Scale and Shape. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 54, n. 3, p. 507–554, abr. 2005. ISSN 0035-9254. DOI: 10.1111/j.1467-9876.2005.00510.x. eprint: https://academic.oup.com/jrsssc/article-pdf/54/3/507/50016951/jrsssc_54_3_507.pdf. Disponível em: <https://doi.org/10.1111/j.1467-9876.2005.00510.x>.

RIGBY, R. A.; STASINOPOULOS, D. M.; HELLER, G.; DE BASTIANI, F. **Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R**. 1st. [S.l.]: Chapman e Hall/CRC, set. 2019. ISBN 9780367278847. DOI: 10.1201/9780429298547.

SERINALDI, F. Distributional modeling and short-term forecasting of electricity prices by Generalized Additive Models for Location, Scale and Shape. **Energy Economics**, v. 33, n. 6, p. 1216–1226, 2011. ISSN 0140-9883. DOI: https://doi.org/10.1016/j.eneco.2011.05.001. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0140988311001058>.

TAYLOR, J.; VERBYLA, A. Joint modelling of location and scale parameters of the t distribution. **Statistical Modelling - STAT MODEL**, v. 4, p. 91–112, jul. 2004. DOI: 10.1191/1471082X04st068oa.

WORLD HEALTH ORGANIZATION. **WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development**. 3. ed. [S.l.]: World Health Organization, 2006. ISBN 924154693X. Disponível em: <https://www.who.int/publications/i/item/924154693X>.

WRIGHT, E.; ROYSTON, P. A Comparison of Statistical Methods for Age-related Reference Intervals. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, v. 160, p. 47–69, jun. 2008. DOI: 10.1111/1467-985X.00045.

APÊNDICE A – CÓDIGOS DA SEÇÃO 3.2

```
library(gamlss)

n <- 50 # definindo tamanho da amostra
lambda = -0.5 # definindo o valor da assimetria
Betas <- c(1, 0.7, 0.5) # betas verdadeiros
Gamas <- c(-0.5, 0.3, 0.2) # gamas verdadeiros

x1 <- runif(n, -1, 1) # variavel explicativa associada ao beta1
x2 <- runif(n, -1, 1) # variavel explicativa associada ao beta2
x3 <- runif(n, -1, 1) # variavel explicativa associada ao beta3
z1 <- runif(n, -1, 1) # variavel explicativa associada ao gama1
z2 <- runif(n, -1, 1) # variavel explicativa associada ao gama2
z3 <- runif(n, -1, 1) # variavel explicativa associada ao gama3

x <- x1 * Betas[1] + x2 * Betas[2] + x3 * Betas[3] # mu da skew-normal
z <- z1 * Gamas[1] + z2 * Gamas[2] + z3 * Gamas[3] # sigma da skew-normal

SNsample <- gamlss.dist::rSN1( # gerando amostra que segue distribuição Skew-Normal
n, # gerando n amostras
mu = x, # com médias x
sigma =  exp(z), # desvios padrao exp(z)
nu = lambda) # e com o valor da assimetria igual a lambda


modelo <- gamlss(
formula = SNsample ~ 0 + x1 + x2 + x3, # modelando parametro de locacao sem intercepto
sigma.formula = ~ 0 + z1 + z2 + z3, # modelando paramentro de escala sem intercepto
family = SN1, # definindo a familia da distribuição
nu.start = lambda, nu.fix = TRUE, # fixando lambda no valor real
save = TRUE, trace = FALSE,
control = gamlss.control( # critérios de parada do ciclo externo
c.crit = 0.001, # para com diferença de deviance menor que 0.001
n.cyc = 20), # ou 20 interações
i.control = glim.control( # critérios de parada do ciclo interno
cc = 0.001, # para com diferença de deviance menor que 0.001
cyc = 50) # ou 50 interações
)

summary(modelo) # para conferir os coeficientes encontrados
```

APÊNDICE B – SIMULAÇÃO MONTE CARLO da SEÇÃO 3.2

```
n <- 50 # define tamanho da amostra
lambda = -0.5 # define valor de lambda
Betas <- c(1, 0.7, 0.5)
Gamas <- c(-0.5, 0.3, 0.2)
nsimu <- 1000 # número de simulações
betas0Estimados <- data.frame(matrix(0, ncol = nsimu, nrow = 3))
sigma0Estimados <- data.frame(matrix(0, ncol = nsimu, nrow = 3))

i <- 1

while (i <= nsimu) {
  tryCatch({ # caso ocorra algum problema de convergência em alguma iteração
    x1 <- runif(n, -1, 1)
    x2 <- runif(n, -1, 1)
    x3 <- runif(n, -1, 1)
    z1 <- runif(n, -1, 1)
    z2 <- runif(n, -1, 1)
    z3 <- runif(n, -1, 1)

    x <- x1 * Betas[1] + x2 * Betas[2] + x3 * Betas[3]
    z <- z1 * Gamas[1] + z2 * Gamas[2] + z3 * Gamas[3]

    SNsample <- gamlss.dist::rSN1(n,
                                  mu = x,
                                  sigma = exp(z),
                                  nu = lambda)

    modelo <- gamlss(formula = SNsample ~ 0 + x1 + x2 + x3,
                     sigma.formula = ~ 0 + z1 + z2 + z3,
                     family = SN1, nu.start = lambda,
                     nu.fix = TRUE, save = TRUE, trace = FALSE)

    betas0Estimados[,i] <- modelo$mu.coefficients
    sigma0Estimados[,i] <- modelo$sigma.coefficients
    i <- i + 1  # Incrementa o índice de iteração apenas se não houver erro
  }, error = function(e) {
    cat("Erro no loop. Reiniciando a iteração ", i, "\n")
  })

}
```

```
apply(betas0Estimados, 1, mean) # média dos Betas estimados
apply(sigma0Estimados, 1, mean) # média dos Gamas estimados

r1 <- mean(apply(apply(betas0Estimados, 2, function(x) {
  (x - Betas)^2
}), 2, sum)) # EQM dos betas

r2 <- mean(apply(apply(sigma0Estimados, 2, function(x) {
  (x - Gamas)^2
}), 2, sum)) # EQM dos gamas
```

APÊNDICE C – SIMULAÇÃO MONTE CARLO da SEÇÃO 3.3

```
n <- 50
Betas <- c(1, 0.75, 0.5)
Gamas <- c(-0.5, 0.3, 0.2)
Omegas <- c(-0.2, 0.5, 1)
nsimu <- 1000
betas0Estimados <- data.frame(matrix(0, ncol = nsimu, nrow = 3))
sigma0Estimados <- data.frame(matrix(0, ncol = nsimu, nrow = 3))
omegas0Estimados <- data.frame(matrix(0, ncol = nsimu, nrow = 3))

i <- 1

while (i <= nsimu) {
  tryCatch({
    x1 <- runif(n, -1, 1)
    x2 <- runif(n, -1, 1)
    x3 <- runif(n, -1, 1)
    z1 <- runif(n, -1, 1)
    z2 <- runif(n, -1, 1)
    z3 <- runif(n, -1, 1)
    w1 <- runif(n, -1, 1)
    w2 <- runif(n, -1, 1)
    w3 <- runif(n, -1, 1)

    x <- x1 * Betas[1] + x2 * Betas[2] + x3 * Betas[3]
    z <- z1 * Gamas[1] + z2 * Gamas[2] + z3 * Gamas[3]
    w <- w1 * Omegas[1] + w2 * Omegas[2] + w3 * Omegas[3]

    SNsample <- gamlss.dist::rSN1(n,
                                  mu = x,
                                  sigma = exp(z),
                                  nu = w)

    modelo <- gamlss(formula = SNsample ~ 0 + x1 + x2 + x3,
                     sigma.formula = ~ 0 + z1 + z2 + z3,
                     nu.formula = ~ 0 + w1 + w2 + w3,
                     family = SN1,
                     save = TRUE, trace = FALSE)

    betas0Estimados[,i] <- modelo$mu.coefficients
    sigma0Estimados[,i] <- modelo$sigma.coefficients
```

```
    omegas0Estimados[,i] <- modelo$nu.coefficients
    i <- i + 1  # Incrementa o índice de iteração apenas se não houver erro
  }, error = function(e) {
    cat("Erro no loop. Reiniciando a iteração ", i, "\n")
  })

}

apply(betas0Estimados, 1, mean) # média dos Betas estimados
apply(sigma0Estimados, 1, mean) # média dos Gamas estimados
apply(omegas0Estimados, 1, mean) # média dos omegas estimados

r1 <- mean(apply(apply(betas0Estimados, 2, function(x) {
  (x - Betas)^2
}), 2, sum)) # MSE dos betas

r2 <- mean(apply(apply(sigma0Estimados, 2, function(x) {
  (x - Gamas)^2
}), 2, sum)) # MSE dos gamas

r3 <- mean(apply(apply(omegas0Estimados, 2, function(x) {
  (x - Gamas)^2
}), 2, sum)) # MSE dos omegas
```

APÊNDICE D – SNIPPET para a MODELAGEM DOS DADOS

O seguinte *snippet* pode ser utilizado para a modelagem inicial dos dados. Para adicioná-lo na IDE RStudio, acesse: **Tools > Global Options > Code > Edit Snippets** (na parte inferior da janela aberta) e cole o *snippet* abaixo. Como a indentação do código não pode ser copiada automaticamente, é necessário indentar manualmente. Para isso, selecione da segunda linha em diante e pressione a tecla **Tab**. Após a inserção correta do *snippet*, basta acessar o script, digitar o nome do *snippet* - neste caso, 'modelagem' - e pressionar **Tab**. A estrutura de todo o código será montada, e a palavra 'dados' aparecerá em negrito. Então, insira o nome do conjunto de dados desejado (por exemplo, 'abdom', utilizado na Seção 4.2). Em seguida, pressione **Tab** novamente para que a palavra 'resposta' fique em negrito e substitua-a pelo nome da variável resposta (como 'y'). Assim, o código inicial estará completo, restando apenas selecionar as variáveis manualmente.

```
snippet modelagem
 library(gamlss)
   ${1:dados}_SN1 <-
   gamlss(${2:resposta} ~ .,
   sigma.formula = ~.,
   nu.formula = ~.,
   data = ${1:dados},
   family = SN1,
   control = gamlss.control(
          n.cyc = 500)  # para maior número máximo de interações do ciclo externo
   )

   plot(${1:dados}_SN1) # gráfico dos resíduos
   wp(${1:dados}_SN1) # worm plot
   wp(${1:dados}_SN1, ylim.all = 1.5) # worm plot com limite de -1.5 a 1.5
   summary(${1:dados}_SN1) # para ver a significância das variáveis




   ${1:dados}_TF <-
   gamlss(${2:resposta} ~ .,
   sigma.formula = ~.,
   tau.formula = ~.,
   data = ${1:dados},
   family = TF,
   control = gamlss.control(
          n.cyc = 500)  # para maior número máximo de interações do ciclo externo
   )
```

```
plot(${1:dados}_TF) # gráfico dos resíduos
wp(${1:dados}_TF) # worm plot
wp(${1:dados}_TF, ylim.all = 1.5) # worm plot com limite de -1.5 a 1.5
summary(${1:dados}_TF) # para ver a significância das variáveis



${1:dados}_NO <-
gamlss(${2:resposta} ~ .,
sigma.formula = ~.,
data = ${1:dados},
family = NO,
control = gamlss.control(
       n.cyc = 500)  # para maior número máximo de interações do ciclo externo
)

plot(${1:dados}_NO) # gráfico dos resíduos
wp(${1:dados}_NO) # worm plot
wp(${1:dados}_NO, ylim.all = 1.5) # worm plot com limite de -1.5 a 1.5
summary(${1:dados}_NO) # para ver a significância das variáveis



${1:dados}_ST2 <-
gamlss(${2:resposta} ~ .,
sigma.formula = ~.,
nu.formula = ~.,
tau.formula = ~.,
data = ${1:dados},
family = ST2,
control = gamlss.control(
       n.cyc = 500)  # para maior número máximo de interações do ciclo externo
)

plot(${1:dados}_ST2) # gráfico dos resíduos
wp(${1:dados}_ST2) # worm plot
wp(${1:dados}_ST2, ylim.all = 1.5) # worm plot com limite de -1.5 a 1.5
summary(${1:dados}_ST2) # para ver a significância das variáveis
```

```
   # seleção dos modelos por AIC
   gamlss::GAIC(${1:dados}_SN1, ${1:dados}_ST2,${1:dados}_TF, ${1:dados}_NO,
   k=2, c = FALSE
)

   # seleção dos modelos por BIC
   gamlss::GAIC(${1:dados}_SN1, ${1:dados}_ST2,${1:dados}_TF, ${1:dados}_NO,
   k=log(nrow(${1:dados})), c = FALSE
)
```

ANEXO A – Estimativas e erros padrão dos parâmetros dos modelos finais.

.1 Martin Marietta

Momento modelado	Parâmetro	Estimativa	Erro padrão
μ_i	Intercepto	-0.05617	0.0112
$\log(\sigma_i^2)$	Intercepto	-2.46023	0.099
	CRSP	12.94994	2.133
λ_i	Intercepto	1.7730	0.769
	CRSP	54.7783	19.86

Tabela 11 – Parâmetros, estimativas e erros padrão do modelo ajustado para o conjunto de dados da Seção 4.1 sob distribuição skew-normal.

.2 Circunferência abdominal

Momento modelado	Parâmetro	Estimativa	Erro padrão
μ_i	Intercepto	-78.36	1.701
	Idade	10.95	0.086
$\log(\sigma_i^2)$	Intercepto	2.18	0.129
	Idade	0.02	0.004
λ_i	Intercepto	5.28	0.788
	Idade	-0.14	0.021

Tabela 12 – Parâmetros, estimativas e erros padrão do ajustado para o conjunto de dados da Seção 4.2 sob distribuição skew-normal.

.3 Produção de leite

Momento modelado	Parâmetro	Estimativa	Erro padrão
μ_i	Intercepto	-1.474×10^5	510.5
	I(lactacao$_i = 2$)	1.386×10^3	42.92
	I(lactacao$_i = 3$)	1.737×10^3	44.34
	Ano do Parto	77.3	2.54
$\log(\sigma_i^2)$	Intercepto	-9.24	2.67
	Idade	0.02	0.007
	I(lactacao$_i = 2$)	0.103	0.036
	I(lactacao$_i = 3$)	0.081	0.058
	Ano do Parto	0.008	0.0013

Tabela 13 – Parâmetros, estimativas e erros padrão do modelo ajustado para o conjunto de dados da Seção 4.3 sob distribuição normal.