FERNANDA VENTURATO ROQUIM

GAMLSS NA EXPERIMENTAÇÃO AGROPECUÁRIA:

UM ESTUDO EM DOENÇAS PARASITÁRIAS DE BOVINOS DE LEITE

LAVRAS – MG

2018

FERNANDA VENTURATO ROQUIM

GAMLSS NA EXPERIMENTAÇÃO AGROPECUÁRIA:
UM ESTUDO EM DOENÇAS PARASITÁRIAS DE BOVINOS DE LEITE

Dissertação apresentada à Universidade Federal de Lavras como parte dos requisitos do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária para obtenção do título de mestre. Área de concentração: Análise de Regressão.

Prof. DSc. Renato Ribeiro de Lima
Orientador

Prof. DSc. Luiz Ricardo Nakamura
Coorientador

LAVRAS – MG

2018

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA.

Roquim, Fernanda Venturato.
GAMLSS na experimentação agropecuária : um estudo em doenças parasitárias de bovinos de leite / Fernanda Venturato Roquim. – Lavras : UFLA, 2018.
82 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2018.
Orientador: Prof. DSc. Renato Ribeiro de Lima.
Bibliografia.

1. Análise de Regressão. 2. GAMLSS. 3. Bezerras Leiteiras. I. de Lima, Renato Ribeiro. II. Nakamura, Luis Ricardo. III. Título.

FERNANDA VENTURATO ROQUIM

GAMLSS NA EXPERIMENTAÇÃO AGROPECUÁRIA: UM ESTUDO EM DOENÇAS PARASITÁRIAS DE BOVINOS DE LEITE
GAMLSS IN AGRICULTURAL AND LIVESTOCK EXPERIMENTATION: A STUDY ON PARASITIC DISEASES OF DAIRY CATTLE

Dissertação apresentada à Universidade Federal de Lavras como parte dos requisitos do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária para obtenção do título de mestre. Área de concentração: Análise de Regressão.

APROVADA em 5 de Junho de 2018.

Profa. DSc. Izabela Regina Cardoso Oliveira	UFLA
Prof. DSc. Marcelo Ângelo Cirillo	UFLA
Prof. DSc. Leandro Ferreira	UNIFAL

Prof. DSc. Renato Ribeiro de Lima
Orientador

Prof. DSc. Luiz Ricardo Nakamura
Co-Orientador

LAVRAS – MG
2018

AGRADECIMENTOS

Em primeiro lugar agradeço ao professor Renato Ribeiro de Lima, por confiar à mim uma pesquisa tão bacana e interessante. Mesmo diante dos inúmeros obstáculos que se colocaram, o professor Renato estava presente, na sua sala cheia de papéis e mesmo super atarefado, conseguia um tempinho para puxar a minha orelha, quando necessário, e também me motivar e orientar. O professor Renato acreditou que eu seria capaz de realizar esta pesquisa, muitas vezes, mais do que eu mesma, e isso foi fundamental. Agradeço, professor, pela confiança depositada na realização deste trabalho.

Agradeço ao professor Luiz Ricardo Nakamura, por todo o conhecimento transmitido sobre os GAMLSS. Mesmo à distância, o professor Nakamura sempre se mostrou disposto à me orientar e tirar dúvidas. Agradeço por todas as contribuições que foram feitas, elas foram muito relevantes e de suma importância para que esta pesquisa acontecesse. Professor, obrigada por todos ensinamentos, não só teóricos mas também sobre a vida acadêmica, obrigada por me compreender quando eu não estava bem e por também confiar que eu conseguiria realizar este trabalho.

Professores Renato e Nakamura, sou feliz por vocês terem aceitado me orientar, e mais ainda, por saber que continuaremos trabalhando juntos.

Agradeço aos professores Marcelo Ângelo Cirillo e Leandro Ferreira por todas as contribuições feitas no momento da defesa. Vocês fizeram contribuições essenciais para melhorar este trabalho, tanto em questões teóricas, quanto aperfeiçoamento da redação. Agradeço também à professora Izabela Regina Cardoso Oliveira pela proposta de um título que melhor representaria esta pesquisa. Muito obrigada por todas as sugestões feitas, foram muito pertinentes.

Agradeço ao meu companheiro Rossi Henrique Soares Chaves, por ser ouvido quando eu era só raiva e ansiedade. Você é um exemplo de ser humano. Obrigada por partilhar sua vida comigo e me transmitir tanta sabedoria. Sou extremamente grata por ter você como namorado, mesmo com a distância sendo dura com nós dois, você sempre se fez presente em todos momento da minha vida. Agradeço por ter você para me ajudar, refletir sobre as coisas e a me fazer ser uma pessoa melhor. Você foi um alicerce fundamental para que eu pudesse realizar esta pesquisa.

Agradeço também minha mãe, Luciene Venturato Roquim, e meu pai, Adauto Roquim, por me darem a dádiva da vida, por sempre estarem ao meu lado, me incentivando e me apoiando incondicionalmente. Sou grata por toda a compreensão que tiveram com o meu processo e peço desculpas pelas vezes em que não pude estar junto de vocês mesmo quando a saudade existia. Eu amo vocês.

Agradeço à Jussara, Pamella, Karol e meu irmão, André, por transformarem a minha casa em um lar. Agradeço também à Ariana e Jorge, por toda a amizade e companheirismo. O carinho que sempre recebi de vocês aqui em Lavras foi muito importante pra mim durante o meu processo do mestrado.

Agradeço à Universidade Federal de Lavras pela oportunidade de estar adquirindo o título de mestra em Estatística e Experimentação Agropecuária e também pelo ambiente amigável e criativo que proporciona para seus alunos, professores e servidores se desenvolverem.

Por último, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES por financiar esta pesquisa.

Não somos o que sabemos.
Somos o que estamos dispostos a aprender.
(Autor Desconhecido)

RESUMO

Neste trabalho foram expostas as principais características dos modelos aditivos generalizados para locação, escala e forma (GAMLSS), apresentando as técnicas de estimação que inclui uma breve descrição dos algoritmos iterativos, as técnicas de inferência e predição, trazendo brevemente sobre as distribuições e termos aditivos possíveis e, por fim, as ferramentas de seleção de modelos e diagnósticos de resíduos. A ênfase foi apresentar a flexibilidade de tal classe de modelos, que nos permite ajustar todos os parâmetros de uma distribuição em função de variáveis explicativas. Em GAMLSS, temos um amplo leque de distribuições para a variável resposta com a possibilidade de incorporar funções de suavização, oferecendo modelos flexíveis que são capazes de descrever bem a realidade. O objetivo deste trabalho consistiu em demonstrar como esta modelagem pode ser utilizada para dados de experimentação agropecuária, trazendo uma aplicação prática ao estudo de tratamentos para parasitoses em bezerras leiteiras. Mostramos também alguns artigos disponíveis no meio científico que já utilizaram os GAMLSS dentro desta área.

Palavras-chave: Bezerras leiteiras. Dados de contagem. Excesso de zeros. Experimentação agropecuária. GAMLSS.

ABSTRACT

This study shows the main characteristics of the generalized additive models for location, scale and shape (GAMLSS). Furthermore, describes its estimation methods, which includes the iterative algorithms, the inference and prediction methods, briefly discussing possible distributions and additive terms and, finally, model selection and diagnostics techniques. The emphasis was to present the flexibility of such class of models which allows us to fit all parameters of the response variable distribution as a function of explanatory variables. In GAMLSS, a wide range of distributions are possible to explain the response variable with the possibility of incorporating smoothing functions, which offer models that are able to increasingly portray reality. The aim of this study was to demonstrate how this models can be used in agricultural and livestock experimental data, bringing a practical application related to the study of treatments for parasitosis in dairy cattle. In addition, it was presented some papers available in the scientific community that have already used the GAMLSS within this area.

Keywords: Agricultural and livestock experimentation. Count data. Dairy cattle. GAMLSS. Inflated zeros.

SUMÁRIO

1 INTRODUÇÃO

Os modelos aditivos generalizados para locação, escala e forma (*generalized additive models for location, scale and shape* – GAMLSS), propostos por Rigby e Stasinopoulos (2005), talvez sejam a técnica de modelagem de regressão mais versátil, disponível no meio científico atualmente. Eles foram propostos com o objetivo de proporcionar modelos com pressupostos mais flexíveis, que, consequentemente, unificam em uma mesma metodologia uma série de metodologias pré-existentes. Esta metodologia aborda, em conjunto, técnicas de modelos lineares e não-lineares, modelos lineares generalizados (GLM) (NELDER; WEDDERBURN, 1972), modelos aditivos generalizados (GAM) (HASTIE; TIBSHIRANI, 1990) e modelos mistos.

A ideia central de um modelo de regressão linear é sugerir uma maneira de se compreender e quantificar a relação de dependência entre uma variável resposta em função de variáveis explicativas, objetivando a obtenção de informações e estudo do comportamento de determinado fenômeno. Entretanto, os primeiros modelos de regressão linear, também chamados de modelos de regressão clássicos, foram criados com pressupostos muito rígidos, que, apesar de simplificarem o modelo, acabam por torná-lo inadequado para diversas situações. Esta modelagem pressupõe que a variável em estudo necessariamente precisa ter distribuição normal com variância constante e que as observações das variáveis em questão sejam independentes entre si, condições estas que nem sempre condizem com a realidade.

Para contornar este problema, algumas alternativas foram criadas, como, por exemplo, a aplicação de uma transformação diretamente na variável resposta em busca de normalizá-la e estabilizar a variância, porém nem sempre este método pode ser aplicado, podendo ainda comprometer a interpretação do modelo. Neste contexto, Nelder e Wedderburn (1972) propuseram os GLM que, podemos assim considerar, aplica uma transformação na esperança da variável resposta, não mais diretamente na variável, ligando o preditor linear ao componente aleatório por meio de uma função de ligação. Esta técnica nos permite ajustar modelos de regressão para variáveis discretas e contínuas, possibilitando a análise de dados binários, categóricos, de contagem, de proporções por meio de distribuições que precisam pertencer à família exponencial.

Apesar de ser amplamente utilizado e de trazer grandes avanços para a análise de regressão, os GLM também podem ser inadequados para algumas situações como, por exemplo, quando o relacionamento entre a média da variável resposta e as variáveis explicativas não é linear. Neste contexto, surgiram os GAM, propostos por Hastie e Tibshirani (1990), que adicionam funções de suavização aos GLM, relaxando esta necessidade, melhorando o ajuste e resolvendo inadequações para alguns casos.

Entretanto, ainda existem casos que não se ajustam à estas modelagens, como por exemplo, quando as distribuições da família exponencial não se adequam, ou quando mais parâmetros da distribuição podem ser melhor explicados se também modelados em função de variáveis explicativas. Assim, os GAMLSS foram propostos, como já mencionado, por Rigby e Stasinopoulos (2005), permitindo o ajuste de modelos que aceitam diversas distribuições para a variável resposta, independente de pertencer à família exponencial, modelos que podem incorporar funções não-paramétricas e/ou efeitos aleatórios, e ainda, não só o parâmetro de locação é modelado, mas também todo e qualquer parâmetro da distribuição. Em síntese, os GAMLSS são um modelo mais aprimorado e flexível, que, além de propor uma nova classe de modelos, também abarca todas as demais citadas, tendo como pressuposto apenas que as observações das variáveis sejam independentes.

Estes modelos têm sido utilizados em diversas áreas, incluindo: atuária, biologia, economia, genômica, finanças, oceanologia, nutrição, medicina, meteorologia, entre outros (STASINOPOULOS et al., 2017). Também inclui-se nesta lista a experimentação agropecuária, um importante setor da economia brasileira, que será a área enfatizada nesta pesquisa.

O desenvolvimento agropecuário está estritamente ligado ao estudo de experimentos, que podem ser realizados em laboratórios ou fazendas experimentais e são, geralmente, analisados por meio de metodologias estatísticas, com intuito de testar hipóteses, propor novas técnicas de cultivo e manejo de animais, por exemplo.

Desta forma, podemos destacar que esta área tem se beneficiado com o uso dos GAMLSS, na medida que já existem alguns estudos publicados sobre o tema, e pode continuar se beneficiando consideravelmente, uma vez que esta é uma modelagem recente e ainda há bastante o que se desenvolver, principalmente no Brasil, visando ganhos econômicos e acadêmicos.

Todo este contexto nos leva ao objetivo principal desta pesquisa, que foi apresentar e discorrer sobre os GAMLSS e mostrar como eles podem ser aplicados à experimentação agropecuária, trazendo uma introdução às técnicas estatísticas desta modelagem, alguns trabalhos publicados e uma aplicação inédita, apresentando o uso do *software* R. Na aplicação, quando analisamos o efeito de dois diferentes tipos de tratamentos para doenças parasitárias em bezerras leiteras, mostramos que os GAMLSS são mais adequados para a natureza dos dados, que apresentam curtose muito elevada e excesso de zeros. Natureza esta, que seria pobremente ajustada por modelagens mais limitadas.

Também, um objetivo secundário foi realizar um trabalho que facilitasse ao usuário compreender o uso de tais modelos. De acordo com a plataforma Google Acadêmico[1], o trabalho de Rigby e Stasinopoulos (2005) já foi citado em 1227 trabalhos e o livro de Stasinopoulos et al. (2017) citado 33 vezes. Enquanto que, o livro de McCullagh e Nelder (1989) já foi utilizado por 33848 trabalhos e o livro de Hastie (2017) já foi citado 15373 vezes. É claro que não podemos desconsiderar que os GAMLSS são uma modelagem bastante recente se comparada aos GLM e GAM, entretanto, podemos constatar que, de maneira geral, a utilização dos GAMLSS ainda é bastante baixa. Assim, esperamos que este trabalho colabore para a disseminação desta metodologia no meio científico nacional.

[1] Disponível em: https://scholar.google.com.br/. Acesso em: 20 de maio de 2018.

2 REFERENCIAL TEÓRICO

Neste capítulo descreveremos brevemente as metodologias de análise de regressão clássica, GLM e GAM, de forma a criar um contexto inicial para compreensão dos GAMLSS e entender como eles foram desenvolvidos. Em seguida, apresentamos os métodos de estimação, inferências, seleção de modelos e diagnósticos de resíduos em GAMLSS. Posteriormente, também exibimos os pacotes e funções do R que podem ser utilizados nas análises e por fim, algumas aplicações encontradas no meio científico relacionadas à experimentação agropecuária.

2.1 Modelos de Regressão

O intuito de um modelo é estabelecer uma relação de dependência quantificável entre variáveis que pode ser expressa por meio de um modelo matemático, que possui todos os seus componentes fixos, ou ainda, por um modelo estatístico, quando incluímos pelo menos um componente aleatório, que é um termo que acrescenta uma variabilidade aleatória ao modelo e é associado à uma distribuição de probabilidade.

Uma classe específica de modelos estatísticos, chamados de modelos de regressão linear (MRL), primeiramente assim nomeados por Francis Galton em 1885 (RODGERS; NICEWANDER, 1988), podem ser definidos por (CHARNET et al., 2008)

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i,$$

em que Y_i são variáveis aleatórias atribuídas às variáveis respostas, com $i = 1, \ldots, n$, β_p são os parâmetros do modelo, com $p = 0, \ldots, k$, x_{ik} são as k variáveis explicativas, ϵ_i são os resíduos, sendo que, $\epsilon_i \overset{iid}{\sim} N(0; \sigma^2)$ e n é o tamanho amostral. Tal modelo assume que os termos dos erros, ϵ_i, são independentes e identicamente distribuídos segundo uma normal com média zero e variância constante σ^2. Isto é equivalente a dizer que

$$Y_i \overset{iid}{\sim} N(\mu_i; \sigma^2),$$

sendo

$$\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}.$$

Estes modelos também podem ser escritos na forma matricial, ou seja,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

e, assim, $\boldsymbol{Y} \overset{iid}{\sim} N(\boldsymbol{\mu}; \boldsymbol{I}\sigma^2)$ e $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, em que $\boldsymbol{Y}$ é o vetor das variáveis respostas de dimensão $n \times 1$, $\boldsymbol{X}$ é a matriz de valores da(s) variável(is) explicativa(s), também chamada de matriz de delineamento, com dimensão $n \times (k+1)$, $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos a ser estimado pelos dados, de dimensão $(k+1) \times 1$, $\boldsymbol{\epsilon}$ é o vetor de valores aleatórios ou vetor dos erros, com dimensão $n \times 1$, em que $\boldsymbol{\epsilon} \overset{iid}{\sim} N(\boldsymbol{0}; \boldsymbol{I}\sigma^2)$ e $\boldsymbol{\mu}$ é o vetor de médias, de dimensão $n \times 1$ (RENCHER; SCHAALJE, 2008).

Se a matriz de delineamento $\boldsymbol{X}$ é de posto completo, o que implica em uma solução única do sistema de equações normais, então trata-se de um modelo de regressão linear. Mas ao tratar de delineamentos experimentais, a matriz $\boldsymbol{X}$ será de posto incompleto. Assim, devemos nos atentar à correta especificação da matriz de parâmetros e alguma restrição deve ser utilizada (SEARLE; CASELLA; MCCULLOCH, 2009). O termo linear refere-se à relação dos parâmetros com a esperança da variável resposta, que, pressupõe-se, ser linear. Se $\boldsymbol{X}$ é de posto incompleto então $(\boldsymbol{X'X})$ não possui inversa simples, e pode-se utilizar a teoria de matrizes inversas generalizadas que aplicam restrições para solucionar o problema (RENCHER; SCHAALJE, 2008).

Considerar que a variável resposta tem distribuição normal é um pressuposto muito forte que exclui uma gama de situações, como, por exemplo, variáveis respostas binárias, de proporção, de contagem ou ainda categóricas. Apesar de ter sido bastante utilizado pela comunidade acadêmica, principalmente na maior parte do século passado (STASINOPOULOS et al., 2017), e ainda ser utilizado até hoje, os MRL são inapropriados para uma série de situações.

Neste contexto, Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (GLM) cuja ideia básica consiste em abrir um leque de opções para a distribuição da variável resposta para além da normal, e, assim, tornando-se uma flexível ampliação dos MRL.

Basicamente, os GLM relacionam a variável resposta com as variáveis explicativas por meio de uma função de ligação, que tem por objetivo garantir que os parâmetros possam assumir quaisquer valores reais preservando os valores ajustados no domínino da distribuição. Ele unifica metodologias de vários modelos estatísticos e, principalmente, considera as distribuições da família exponencial. Assim, denotamos $\boldsymbol{Y} \sim \mathcal{FE}(\boldsymbol{\mu}, \boldsymbol{\phi})$, em que $\boldsymbol{\mu}$ e $\boldsymbol{\phi}$ são vetores dos parâmetros de locação e escala da distribuição, respectivamente, sendo que $\boldsymbol{\phi}$ é constante. Distribuições populares, como a normal, Poisson, binomial, binomial negativa, gama, beta normal inversa, são exemplos de distribuições que pertencem à família exponencial e podem ser usadas no ajuste de GLM.

Em GLM, a variável aleatória $\boldsymbol{Y}$ é relacionada à uma função de probabilidade (fp) ou uma função densidade de probabilidade (fdp) e é chamada de componente aleatório. Portanto, seja $\boldsymbol{\mu}$ a média da distribuição, tem-se que

$$E(\boldsymbol{Y}) = \boldsymbol{\mu}.$$

As variáveis explicativas ($\boldsymbol{X}$) e os parâmetros do modelo ($\boldsymbol{\beta}$) formam o componente sistemático. Esse componente é chamado de preditor linear ($\boldsymbol{\eta}$), sendo que $\boldsymbol{\eta}$ é um vetor com dimensão n, portante, temos que

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

O terceiro componente de um GLM é a função de ligação, que tem por objetivo ligar o componente aleatório ao componente sistemático, ou seja, relaciona a média ao preditor linear ($g(\boldsymbol{\mu}) = \boldsymbol{\eta}$). As funções de ligação ($g(\cdot)$) precisam necessariamente ser diferenciáveis e monótonas, dadas por,

$$g(E[\boldsymbol{Y}]) = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

(DEMÉTRIO, 2001). Destacamos aqui três funções de ligação comumente utilizadas: identidade, dada por $\boldsymbol{\eta} = \boldsymbol{\mu}$, que, se utilizada em GLM com resposta normal resulta no MRL; logarítmica, dada por $\boldsymbol{\eta} = ln(\boldsymbol{\mu})$, comumente utilizada para respostas de contagem; e logística, dada por $\boldsymbol{\eta} = ln\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right)$, usada para dados binários.

Devemos fazer uma análise descritiva inicial da variável resposta, para realizar a melhor escolha da distribuição a ser utilizada, como identificar a natureza discreta ou

contínua, características de simetria e intervalo em que os valores são observados. Assim, escolhe-se a função de ligação que seja compatível com a distribuição a ser utilizada, e que garanta, quando possível, boa interpretação para o modelo. Por fim, formula-se o componente sistemático do modelo, que pode conter variáveis quantitativas ou fatores.

Em suma, os GLM trazem duas principais inovações em relação ao MRL que são importantes na compreensão dos GAMLSS. A primeira é a flexibilização da distribuição da variável resposta, que agora engloba não só a normal, mas qualquer distribuição que pertença à família exponencial. A segunda é a utilização de uma função de ligação para modelar a relação entre $E(\boldsymbol{Y})$ e as variáveis explicativas.

Apesar de flexibilizar o pressuposto de normalidade dos MRL, os GLM ainda mostram-se limitados na medida em que considera que a relação da média é linear às variáveis explicativas. Hastie e Tibshirani (1990) introduziram a técnica de suavização aos GLM, dando origem aos chamados modelos aditivos generalizados (GAM). Os GAM foram pensados como uma alternativa para melhorar o ajuste dos GLM, adicionando uma função de suavização não-paramétrica nas covariáveis de forma a deixar os próprios dados conduzirem a sua relação com o preditor (η), que muitas vezes, acontecem de forma não-linear. Os GAM podem ainda, eventualmente, resolver problemas de resíduos assimétricos e/ou heterocedásticos, por exemplo. Um GAM pode ser escrito como,

$$\begin{gathered}\boldsymbol{Y} \overset{ind}{\sim} \mathcal{FE}(\boldsymbol{\mu}, \boldsymbol{\phi}) \\ \boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + s_1(\boldsymbol{x}_1) + \ldots + s_J(\boldsymbol{x}_J)\end{gathered}^{,}$$

em que s_j é uma função de suavização não-paramétrica aplicada à covariável $\boldsymbol{x}_j$ para $j = 1, \ldots, J$ (STASINOPOULOS et al., 2017). Observe que nem todas as covariáveis precisam receber funções de suavização. Podemos citar aqui algumas funções de suavização mais conhecidas, como *P-splines*, splines cúbicas, *loess* e redes neurais. Uma spline é uma curva definida matematicamente por dois ou mais pontos de controle e é dessa forma que deixamos os dados determinarem a relação entre $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ e as variáveis explicativas, que é a ideia central de se usar uma função de suavização (EILERS; MARX; DURBÁN, 2015).

Quando utilizarmos um GAM, devemos tomar cuidado ao realizar a análise do modelo, porque os erros padrões são inválidos na medida que só podem ser calculados com base na covariável em si, pressupondo uma relação linear, que não leva em consideração o efeito das funções de suavização como um todo.

Uma das características do ajuste de modelos semi-paramétricos é que eles não podem ser facilmente descritos em uma forma matemática, entretanto, eles podem ser exibidos de forma gráfica, que conjuntamente pode nos trazer uma série de informações sobre a adequação do ajuste e resíduos.

Os GLM e os GAM ainda são limitados, uma vez que apenas o parâmetro de locação (média) das distribuições é modelado, e necessariamente, a distribuição precisa pertencer à família exponencial. Existem situações em que podemos requerer mais flexibilidade para a distribuição da variável resposta.

Neste contexto, Rigby e Stasinopoulos (2005) propuseram a classe de modelos de regressão GAMLSS, permitindo o ajuste de modelos que aceitam qualquer distribuição para a variável resposta, independente dela pertencer à alguma família de distribuições. Também em GAMLSS, a parte sistemática do modelo é expandida, de forma que, não só a média da distribuição é modelada, mas todos os parâmetros de locação, escala e forma da distribuição da variável resposta. Assim, todos esses parâmetros podem ser modelados em função das variáveis explicativas e, além disso, os preditores também podem incorporar funções não-paramétricas de suavização, efeitos aleatórios, ou outros termos aditivos. Além de propor uma nova classe de modelos, os GAMLSS possuem os modelos MRL, GLM e GAM como casos particulares. Porém estes modelos ainda assumem que as observações da variável resposta Y são independentes entre si.

Um modelo GAMLSS pode ser definido por

$$
\begin{aligned}
\boldsymbol{Y} &\overset{ind}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + s_{11}(\boldsymbol{x}_{11}) + \ldots + s_{1J_1}(\boldsymbol{x}_{1J_1}) \\
\boldsymbol{\eta}_2 = g_2(\boldsymbol{\sigma}) &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + s_{21}(\boldsymbol{x}_{21}) + \ldots + s_{2J_2}(\boldsymbol{x}_{2J_2}) \\
\boldsymbol{\eta}_3 = g_3(\boldsymbol{\nu}) &= \boldsymbol{X}_3\boldsymbol{\beta}_3 + s_{31}(\boldsymbol{x}_{31}) + \ldots + s_{3J_3}(\boldsymbol{x}_{3J_3}) \\
\boldsymbol{\eta}_4 = g_4(\boldsymbol{\tau}) &= \boldsymbol{X}_4\boldsymbol{\beta}_4 + s_{41}(\boldsymbol{x}_{41}) + \ldots + s_{4J_4}(\boldsymbol{x}_{4J_4}),
\end{aligned}
\tag{2.1}
$$

em que $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ é uma distribuição de quatro parâmetros, $\boldsymbol{\mu}$ usualmente é um parâmetro de locação, $\boldsymbol{\sigma}$ é, frequentemente, um parâmetro de escala, $\boldsymbol{\nu}$ e $\boldsymbol{\tau}$ são os parâmetros de forma da distribuição, geralmente associados à assimetria e curtose, respectivamente. As matrizes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ e $\boldsymbol{X}_4$ podem ou não coincidir, isto é, o preditor de cada parâmetro da distribuição pode receber diferentes variáveis explicativas (RIGBY; STASINOPOULOS, 2005).

O modelo descrito na Equação 2.1 é a definição do GAMLSS com a maior quantidade de parâmetros que temos implementado hoje no *software* R (R Core Team, 2017), entretanto, teoricamente, também é possível assumir uma distribuição de probabilidade para a variável resposta com mais de quatro parâmetros, incluindo mais preditores em tal definição. Ainda poderíamos definir modelos intermediários para distribuições com dois ou três parâmetros e ajustar σ e/ou ν em função de covariáveis, bastando reduzir a quantidade de preditores.

É importante ressaltarmos, que cada parâmetro pode receber uma função de ligação diferente. A função de ligação utilizada em cada preditor, como proposto por Rigby e Stasinopoulos (2005), está relacionada à amplitude dos valores dos parâmetros, diferentemente de como era definida em GLM, que a função de ligação estava mais relacionada à distribuição escolhida para a variável resposta. Citando as mesmas funções de ligação, em GAMLSS, a ligação identidade é adotada quando o parâmetro assume valores entre $(-\infty, \infty)$, logarítmica para $(0, \infty)$ e logística para $(0, 1)$.

As distribuições que são permitidas para um GAMLSS são amplas, elas só precisam ser paramétricas. Atualmente, no pacote gamlss (STASINOPOULOS; RIGBY, 2007) do R, existem mais de 100 distribuições implementadas, dentre discretas, contínuas e mistas, incluindo distribuições fortemente assimétricas, platicúrticas ou leptocúrticas. Também é possível implementar uma nova distribuição à critério. Todas as distribuições implementadas no pacote podem ser: truncadas à direita, à esquerda, ou ambos lados; censuradas em qualquer intervalo de resposta; misturadas criando distribuições de mistura; inflacionadas em zero e\ou um. Em relação às distribuições contínuas, estas podem ser discretizadas para modelar respostas discretas. Ainda, distribuições contínuas no intervalo $(-\infty, \infty)$ podem facilmente ser transformadas para intervalos $(0, \infty)$ e $(0, 1)$. Mais detalhes sobre as distribuições podem ser encontrados em Rigby et al. (2017) e no Apêndice B.

As variáveis explicativas podem afetar os parâmetros da distribuição de muitas maneiras. Os modelos GAMLSS podem ser ajustados via funções paramétricas lineares ou não-lineares e funções não-paramétricas. Atualmente, o pacote `gamlss`, isoladamente, permite os seguintes termos aditivos: P-spline (B-spline penalizada); P-spline monótona; P-spline cíclica; P-spline de coeficientes variantes; splines cúbicas; loess; polinômios fracionários; efeitos aleatórios; regressão *ridge* e ajustes não-paramétricos. Em conjunto com

outros pacotes, permite, concomitantemente, o ajuste de: redes neurais; árvores de decisão; efeitos aleatórios e suavizações multidimensionais (STASINOPOULOS et al., 2017).

2.2 Estimação

Nesta seção serão apresentadas algumas técnicas de estimação dos parâmetros e hiperparâmetros dos modelos GAMLSS. Também serão descritos dois algoritmos iterativos de estimação, o algoritmo RS, cuja sigla advém dos autores Rigby e Stasinopoulos (2005), e CG, proposto por Cole e Green (1992).

Os modelos GAMLSS foram previamente definidos na Equação 2.1. Note que grande parte das suavizações podem ser escritas como (STASINOPOULOS et al., 2017)

$$\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{Z}\boldsymbol{\gamma},$$

em que $\boldsymbol{Z}$ é a matriz de base dependendo dos valores de $\boldsymbol{x}$, e $\boldsymbol{\gamma}$ é um conjunto de parâmetros sujeitos à penalização $\boldsymbol{\lambda}\boldsymbol{\gamma}^\top\boldsymbol{G}\boldsymbol{\gamma}$, para uma matriz conhecida $\boldsymbol{G} = \boldsymbol{D}^\top\boldsymbol{D}$ em que $\boldsymbol{D}$ é uma matriz de diferenças e $\boldsymbol{\lambda}$ é um vetor ou escalar de hiperparâmetros que regula o grau de suavização necessário no ajuste.

Desta forma o modelo (2.1) pode ser generalizado e escrito como

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\gamma} &\overset{ind}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_{11}\boldsymbol{\gamma}_{11} + \ldots + \boldsymbol{Z}_{1J_1}\boldsymbol{\gamma}_{1J_1} \\
\boldsymbol{\eta}_2 = g_2(\boldsymbol{\sigma}) &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_{21}\boldsymbol{\gamma}_{21} + \ldots + \boldsymbol{Z}_{2J_2}\boldsymbol{\gamma}_{2J_2} \\
\boldsymbol{\eta}_3 = g_3(\boldsymbol{\nu}) &= \boldsymbol{X}_3\boldsymbol{\beta}_3 + \boldsymbol{Z}_{31}\boldsymbol{\gamma}_{31} + \ldots + \boldsymbol{Z}_{3J_3}\boldsymbol{\gamma}_{3J_3} \\
\boldsymbol{\eta}_4 = g_4(\boldsymbol{\tau}) &= \boldsymbol{X}_4\boldsymbol{\beta}_4 + \boldsymbol{Z}_{41}\boldsymbol{\gamma}_{41} + \ldots + \boldsymbol{Z}_{4J_4}\boldsymbol{\gamma}_{4J_4},
\end{aligned} \tag{2.2}$$

em que os parâmetros de efeito fixo são representados por

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top)^\top$$

e Stasinopoulos et al. (2017) definem

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}^\top, \ldots, \boldsymbol{\gamma}_{1J_1}^\top, \boldsymbol{\gamma}_{21}^\top, \ldots, \boldsymbol{\gamma}_{4J_4}^\top)^\top$$

como os parâmetros de um efeito aleatório assumindo que todos os $\boldsymbol{\gamma}_{kj}$ são independentes entre si e que

$$\boldsymbol{\gamma}_{kj} \sim N\left(0, [\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})]^{-1}\right)$$

sendo $[\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})]^{-1}$ a inversa (generalizada) da matriz simétrica $\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})$ de ordem $q_{kj} \times q_{kj}$.

Se não existem estes efeitos aleatórios no modelo então o modelo 2.2 reduz-se à

$$\begin{aligned}
\boldsymbol{Y} &\overset{ind}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 \\
\boldsymbol{\eta}_2 &= g_2(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 \\
\boldsymbol{\eta}_3 &= g_3(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 \\
\boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4.
\end{aligned} \tag{2.3}$$

Stasinopoulos et al. (2017) denotam 2.3 como modelo GAMLSS paramétrico e 2.2 como modelo GAMLSS com efeito aleatório. Um GAMLSS paramétrico, ou seja, que não tem funções de suavização (efeitos aleatórios definidos acima), requer apenas a estimação de $\boldsymbol{\beta}$. Já um modelo GAMLSS com efeitos aleatórios requer não só a estimação de $\boldsymbol{\beta}$, mas também $\boldsymbol{\gamma}$ e $\boldsymbol{\lambda}$ que é dado por

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{11}^\top, \ldots, \boldsymbol{\lambda}_{1J_1}^\top, \boldsymbol{\lambda}_{21}^\top, \ldots, \boldsymbol{\lambda}_{4J_4}^\top)^\top.$$

Os GAMLSS paramétricos são estimados por máxima verossimilhança, referindo-se à estimação de $\boldsymbol{\beta}$ apenas. Dizer que $\boldsymbol{Y} \overset{ind}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ implica que o logaritmo da função de verossimilhança, definida pela verossimilhança observada da amostra, é

$$l = \sum_{i=1}^{n} ln[f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)] \tag{2.4}$$

sob o pressuposto de independência das observações.

Para um modelo GAMLSS com efeitos aleatórios, é utilizado o método de estimação por máxima verossimilhança penalizada, que refere-se à estimação de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ para $\boldsymbol{\lambda}$ constante. Assim, a função de verossimilhança penalizada é dada por

$$l_p = l - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}^\top \boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})\boldsymbol{\gamma}_{kj}. \tag{2.5}$$

Rigby e Stasinopoulos (2005) sugerem dois algoritmos para ajustar um GAMLSS para valores fixos de hiperparâmetros, a fim de maximizar a função de verossimilhança penalizada, são eles os algoritmos CG e RS.

O primeiro, algoritmo CG, é uma generalização do algoritmo de Cole e Green (1992), originando a sigla CG, que utiliza as primeiras derivadas e os valores exatos ou aproximados das derivadas segundas e derivadas cruzadas da função de verossimilhança dos dados desde que respeitem $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$. Entretanto, para muitas fdp $f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$, os parâmetros possuem informação ortogonal, ou seja, os valores das derivadas cruzadas da função de verossimilhança são iguais à zero. Neste caso, utiliza-se o algoritmo RS, proposto por Rigby e Stasinopoulos (2005), cuja sigla também advém do nome dos autores, que não utiliza o valor esperado das derivadas cruzadas. Ainda, existe uma terceira metodologia, que mistura passos de ambos algoritmos, iniciando o processo por RS e finalizando por CG.

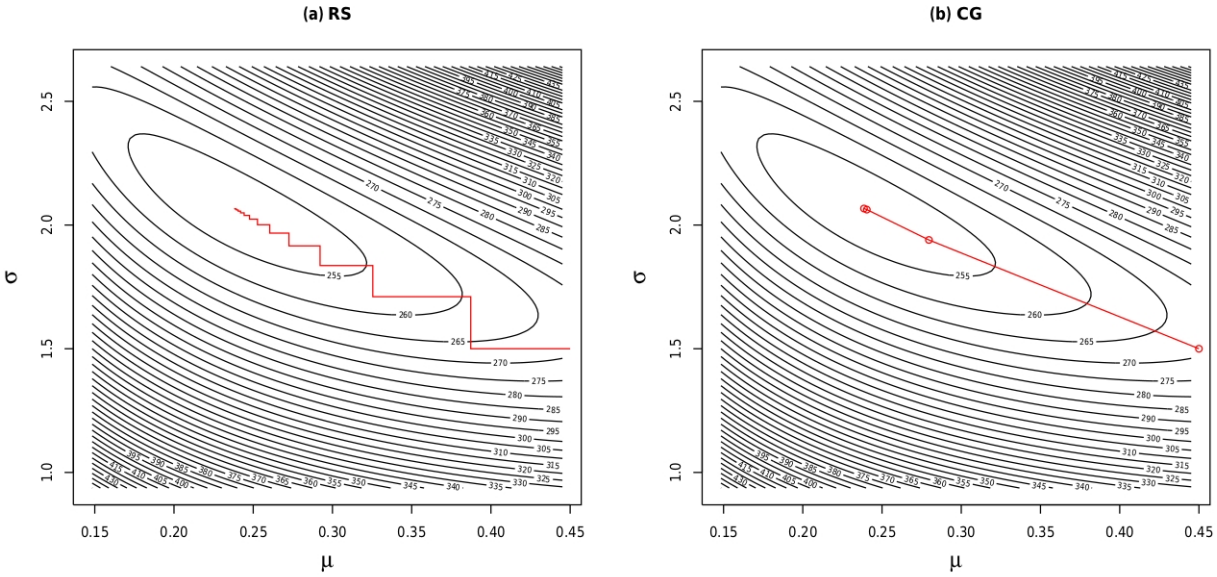
Na Figura 2.1, temos esboçado o caminho que ambos algoritmos fazem para obter o máximo da função de verossimilhança. As elipses representam o desvio global, que será apresentado na Seção 2.4. O importante aqui é destacar que o algoritmo RS maximiza a verossimilhança em cada parâmetro (μ, σ, ν e τ) por vez até atingir a convergência (Gráfico (a)). Já o algoritmo CG tem a capacidade, desde que as derivadas cruzadas não sejam nulas, de atualizar os parâmetros conjuntamente em cada iteração, como mostrado no Gráfico (b). Esta é a principal diferença entre eles. Destacamos também, que em geral, o algoritmo RS é mais estável e mais rápido, se comparado ao CG (STASINOPOULOS et al., 2017).

No trabalho de Rigby e Stasinopoulos (2005), os autores provam que tais algoritmos levam ao máximo da função de verossimilhança penalizada, fornecendo estimadores dos parâmetros de efeitos fixos e aleatórios, $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\gamma}}$ para $\boldsymbol{\lambda}$ constante. Mais detalhes sobre estes algoritmos podem ser encontrados nas seções 3.2.1 e 3.2.2 do livro de Stasinopoulos et al. (2017).

Para obter o máximo da função de verossimilhança penalizada partimos da premissa de que o hiperparâmetro $\boldsymbol{\lambda}$ é constante. Porém, como estimar este grau de suavização?

Os primeiros métodos propostos, chamados de globais porque o método é aplicado fora das iterações dos algoritmos RS e CG, consistiam em utilizar algum procedimento

Figura 2.1 – Processo iterativo gráfico dos algoritmos RS (a) e CG (b) para uma distribuição Weibull(μ, σ) (STASINOPOULOS et al., 2017, p. 63).



baseado em otimização numérica para minimizar o critério de Akaike e assim selecionar a melhor estimativa para $\boldsymbol{\lambda}$, ou, utilizar a minimização do desvio global de um banco de dados de validação para obter os parâmetros de suavização. Entretanto, ambos os métodos exigiam que vários modelos fossem reajustados, o que poderia ser lento computacionalmente, principalmente se o modelo tiver várias funções de suavização (RIGBY; STASINOPOULOS, 2014).

Neste contexto, Rigby e Stasinopoulos (2014) propuseram uma maneira de selecionar este parâmetro de forma automática, que chamam de estimação interna, ou também local, porque o método é aplicado em cada passo do algoritmo RS e CG, utilizando as teorias de máxima verossimilhança. Esta alternativa tem a principal vantagem de ser mais rápida em comparação às demais, e, no trabalho referenciado, os autores mostram que o desempenho é bem semelhante às técnicas anteriores.

Os métodos de estimação dos hiperparâmetros podem ser bastante complexos. Mais informações podem ser encontradas nas seções 3.4.1 e 3.4.3 de Stasinopoulos et al. (2017).

2.3 Inferência

Esta seção tem por objetivo mostrar como são feitas as inferências em GAMLSS, mostrando como podem ser obtidos erros padrões, intervalos de confiança e predições. As inferências aqui apresentadas são aproximações baseadas em verossimilhança.

Para modelos paramétricos, a estimação dos parâmetros é feita por máxima verossimilhança, enquanto que para modelos GAMLSS semi-paramétricos, com termos de suavização, é utilizada máxima verossimilhança penalizada, conforme apresentado na seção anterior.

Para modelos paramétricos, se o modelo foi bem ajustado por meio do método da máxima verossimilhança, então temos que, assintoticamente, todos os estimadores dos parâmetros são consistentes, com erros padrões assintóticos corretos, o que nos leva também à cobertura correta de intervalos de confiança. Se um modelo GAMLSS não é correto, ou seja, ajusta-se pobremente aos dados, então os parâmetros podem não ser consistentes (STASINOPOULOS et al., 2017).

As inferências em modelos de regressão paramétricos, GLM, por exemplo, geralmente se concentram em selecionar ou testar diferentes modelos, estimar os parâmetros, testar valores de um ou mais parâmetros, estimar intervalos de confiança para os parâmetros e realizar predições de valores futuros da variável resposta. Para modelos com funções de suavização, GAM, por exemplo, são necessárias ferramentas extras para realizar inferências que nos permitam tirar conclusões sobre os parâmetros e o comportamento das funções de suavização. Para os GAMLSS, precisamos utilizar ferramentas que sejam capazes de lidar com a diversidade de distribuições e termos no modelo, além de diferentes conjuntos de variáveis explicativas e função de ligação que cada parâmetro do modelo pode receber.

Em GAMLSS, não necessariamente todos os parâmetros da distribuição irão depender de covariáveis, ou seja, podemos observar algum parâmetro constante. Nestes casos, é possível obter erros padrões aproximados e, consequentemente, intervalos de confiança para este parâmetro. Para os casos em que os parâmetros forem ajustados em função de variáveis explicativas, podemos obter os erros padrões e intervalos de confiança para cada coeficiente do preditor linear, desde não seja aplicado funções de suavização naquela covariável. Para fatores, geralmente, não há interesse em analisar os coeficientes, mas sim a contribuição daquele fator no modelo (STASINOPOULOS et al., 2017).

O formato das curvas ajustadas de um termo de suavização e seus referidos erros padrões são muito importantes quando o modelo possui componentes não-paramétricos. Para as funções de suavização mais comuns, os erros padrões das curvas ajustadas costumam ser função do traço da matriz de suavização. A estimação do parâmetro de suavização ($\boldsymbol{\lambda}$) é de suma importância porque determinará o formato da curva de suavização. É raro o interesse em erros padrões deste parâmetro, entretanto, eles podem ser aproximados por técnicas de *bootstrapping*.

Também podemos ter interesse em predizer um valor futuro para qualquer parâmetro da distribuição da variável resposta, seja ele μ, σ, ν, ou τ, ou ainda predizer a própria distribuição ajustada da variável resposta como um todo, que pode ser obtida ao substituir, na distribuição, todos os parâmetros preditos. É preciso ressaltar que, para qualquer que seja a distribuição resultante, não é levado em conta as incertezas dos parâmetros preditos (STASINOPOULOS et al., 2017).

Stasinopoulos et al. (2017) classificam os métodos inferenciais usados para responder as questões acima em duas categorias: inferência baseada em verossimilhança e *bootstrapping*. Aqui abordaremos os métodos baseados em verossimilhança. O autores propõem, inicialmente, um GAMLSS paramétrico (Equação 2.3) e que $\boldsymbol{\theta}$ é o vetor genérico de parâmetros. Neste contexto, pode ser pensado como um conjunto de todos os coeficientes lineares para μ, σ, ν e τ, por exemplo, $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$. Considerando a teoria de máxima verossimilhança clássica, temos que, assintoticamente

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_T, \boldsymbol{i}(\boldsymbol{\theta}_T)^{-1}),$$

em que $\hat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança e

$$\boldsymbol{i}(\boldsymbol{\theta}_T) = -E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T}$$

é a matriz de informação esperada de Fisher assumindo valores verdadeiros de $\boldsymbol{\theta}_T$. Nem sempre é possível derivar a matriz de informação esperada ($\boldsymbol{i}(\boldsymbol{\theta}_T)$) analiticamente. Então, é utilizada a matriz de informação observada de Fisher, definida por

$$\boldsymbol{I}(\boldsymbol{\theta}_T) = -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T}.$$

Note que $\boldsymbol{I}(\boldsymbol{\theta}_T)$ é igual ao negativo da matriz hessiana do logaritmo da verossimilhança em $\boldsymbol{\theta}_T$. A matriz de variâncias e covariâncias da distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é agora aproximada por $\boldsymbol{I}(\boldsymbol{\theta}_T)^{-1}$ ao invés de $\boldsymbol{i}(\boldsymbol{\theta}_T)^{-1}$. Substituí-se $\boldsymbol{\theta}_T$ por $\hat{\boldsymbol{\theta}}$ quando $\boldsymbol{\theta}_T$ é desconhecido, o que nos leva às matrizes de informação esperada e observada, dadas por $\boldsymbol{i}(\hat{\boldsymbol{\theta}})$ e $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$, respectivamente (STASINOPOULOS et al., 2017).

Então, para GAMLSS paramétricos, temos a seguinte distribuição assintótica para $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_T, \boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}).$$

Normalmente, os erros padrões dos parâmetros estimados são obtidos pela raiz quadrada da diagonal da matriz de variâncias e covariâncias associada à θ. Um método alternativo para se obter os erros padrões (EP) de $\hat{\boldsymbol{\beta}}$, quando a matriz de informação é de difícil obtenção, é

$$EP(\hat{\boldsymbol{\beta}}) \approx \frac{|\hat{\boldsymbol{\beta}}|}{\sqrt{\Delta GDEV}},$$

em que $\Delta GDEV$ é a diferença no desvio global obtida por omitir variáveis explicativas associadas à $\boldsymbol{\beta}$. Este resultado se baseia na aproximação da estatística de teste de razão de verossimilhanças à estatística do teste de Wald, sob a hipótese nula de que $\boldsymbol{\beta} = 0$, em que (STASINOPOULOS et al., 2017)

$$\left[\frac{\hat{\boldsymbol{\beta}}}{EP(\hat{\boldsymbol{\beta}})}\right]^2 \approx \Delta GDEV.$$

O desvio global será detalhado na Seção 2.4. Obtido os erros padrões aproximados, podemos obter intervalos de confiança e predições. Mais detalhes sobre a obtenção destas inferências serão mostrados na Seção 2.6. As técnicas de seleção de modelos são apresentadas a seguir.

2.4 Seleção de modelos

As técnicas de seleção de modelos de regressão buscam resolver o problema de se selecionar preditores adequados dentre uma infinidade de possíveis preditores em potencial. A seleção de modelos em GAMLSS envolve a seleção da melhor distribuição para a va-

riável resposta, dos preditores adequados para os parâmetros da distribuição selecionada, das funções de ligação e dos hiperparâmetros.

A avaliação do modelo estatístico geralmente está relacionada à sua capacidade explicativa ou preditiva relativa à um conjunto de dados independentes, geralmente chamado de conjunto de dados teste (STASINOPOULOS et al., 2017). Em geral, reconhece-se que os modelos sobreajustados, ou seja, que possuem interpretações muito complexas, ou subajustados, quando não representam bem os dados, não são muito bons para explicação ou previsão.

O desvio global (GDEV), importante medida para a seleção de modelos em GAMLSS, é definido por

$$GDEV = -2l(\hat{\boldsymbol{\theta}}),$$

em que $l(\hat{\boldsymbol{\theta}})$ é o logarítmo da função de verossimilhança ajustada, apresentado nas Equações 2.4 e 2.5. Esta quantidade é utilizada na definição do critério de Akaike generalizado (GAIC), apresentada em Voudouris et al. (2012), dado por,

$$GAIC(\kappa) = GDEV + (\kappa \times df),$$

em que df denota o total efetivo de graus de liberdade do modelo e κ é a penalidade para cada grau de liberdade utilizado. Se $\kappa = 2$ então o critério coincide com o critério de Akaike (AIC) (AKAIKE, 1998). Se Se $\kappa = ln(n)$ então o critério coincide com o critério de informação bayesiano (BIC)(SCHWARZ et al., 1978). O $GAIC(\kappa)$ penaliza modelos com muitos parâmetros, de forma que, para algum κ escolhido, quanto menor o valor de $GAIC(\kappa)$, melhor ajustado é considerado o modelo (STASINOPOULOS et al., 2017).

Em geral, as inferências sobre um modelo são feitas condicionadas a um único modelo, chamado de modelo final ou melhor modelo. Entretanto devemos tomar cuidado com esta nomenclatura, no sentido de que não existe apenas um único modelo possível de se prever ou explicar bem os dados. Além disso, ao utilizar técnicas de seleção de modelos para definir um modelo estatístico, devemos levar em conta as questões substanciais de interesse e não apenas a teoria matemática e estatística isoladamente. Isso significa que diferentes problemas podem requerer diferentes estratégias de seleção (STASINOPOULOS et al., 2017).

Stasinopoulos et al. (2017) dizem que na busca de um modelo GAMLSS apropriado, para qualquer banco de dados, os componentes relacionados à distribuição, à função de ligação, às possíveis variáveis explicativas a serem inclusas nos preditores e aos hiperparâmetros precisam ser especificados da maneira mais objetiva possível.

A seleção da melhor distribuição pode ocorrer em dois estágios, o de ajuste e o de diagnóstico. O estágio de ajuste envolve a comparação de modelos ajustados de diferentes distribuições usando o $GAIC(\kappa)$. O estágio de diagnóstico envolve o uso de um gráfico bem específico, o *worm plot* (gráfico de minhoca) (BUUREN; FREDRIKS, 2001), que permite a detecção de inadequações no modelo. Trataremos melhor sobre este gráfico na Seção 2.5.

Com relação à escolha da função de ligação, devemos selecioná-la de forma a garantir que tais parâmetros estejam sempre dentro de seus respectivos intervalos, como definido na Seção 2.1. Existem ocasiões em que a escolha da função de ligação é importante do ponto de vista de proporcionar interpretações para os parâmetros. A escolha da função de ligação pode afetar consideravelmente o ajuste do modelo. Diferentes funções de ligação podem ser comparadas usando o $GAIC$ ou o gráfico de minhoca (STASINOPOULOS et al., 2017).

A seleção de variáveis explicativas é, na prática, um dos assuntos mais importantes no ajuste do modelo estatístico. Seja $\mathcal{X}_k$ um conjunto de variáveis explicativas disponíveis para consideração na modelagem do parâmetro $\boldsymbol{\theta}_k$ de um modelo GAMLSS, em que $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4) = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$. Usualmente, $\mathcal{X}_k$ conterá fatores e variáveis quantitativas que podem entrar no modelo de forma linear ou por meio de funções de suavização (STASINOPOULOS et al., 2017).

Stasinopoulos et al. (2017) citam quatro procedimentos disponíveis hoje na literatura estatística para a seleção de variáveis: métodos baseados em critérios, regularização, procedimentos *dimension-reduction* e *boosting*. Enfatizaremos os métodos baseados em critérios, porque são a forma mais utilizada em seleção de variáveis em GAMLSS.

Os procedimentos usuais para seleção de variáveis via critério de informação são os mesmos de um modelo de regressão, podendo-se citar os algoritmos *forward*, *backward* e *stepwise*, que podem ser aplicados para cada parâmetro da distribuição. No procedimento *forward*, cada variável que ainda não está no modelo para aquele parâmetro da distribuição é testada por inclusão. O procedimento é interrompido quando nenhuma das variáveis

restantes é significativa quando adicionada ao modelo, com base no critério de seleção escolhido.

Uma das principais desvantagens do procedimento *forward* é o fato de que cada adição de uma nova variável pode tornar uma ou mais variáveis já inclusas não significativas.

Por outro lado, o procedimento *backward* começa com o ajuste de um modelo com todas as variáveis de interesse. Em seguida, a variável menos significativa é descartada, desde que não seja significativa com base no critério de seleção escolhido. Continuamos a ajustar sucessivamente os modelos reduzidos até que todas as variáveis remanescentes no modelo sejam significativas. Porém, este procedimento tem o mesmo problema do primeiro.

Finalmente, no procedimento *stepwise*, todas as variáveis atualmente no modelo são individualmente consideradas para serem descartadas em cada etapa, enquanto que, todas as variáveis que não estão atualmente no modelo são consideradas para adição. A variável a ser adicionada ou retirada é escolhida naquele passo, desde que reduza o critério. Por este motivo o procedimento *stepwise* é o mais utilizado (STASINOPOULOS et al., 2017).

A quantidade de variáveis explicativas disponíveis em $\mathcal{X}_k$ é muito importante no que diz respeito à seleção das variáveis. Se existem poucas variáveis então temos uma quantidade tratável de possíveis combinações, mas se temos disponíveis muitas covariáveis, o número de todas as possíveis combinações pode rapidamente se tornar muito elevado, de forma a impossibilitar o ajuste de tantos modelos. Por este motivo, é fortemente aconselhável que seja feita uma análise da relação entre as variáveis explicativas, de forma a evitar problemas de colinearidade e selecionar as variáveis explicativas mais relevantes para aquele determinado estudo.

Por último, sobre a seleção dos hiperparâmetros, podemos destacar que os parâmetros de suavização podem ser constantes, definidos pelo pesquisador, ou estimados a partir dos dados conforme visto na Seção 2.2. Um método padrão para definir os parâmetros de suavização, sugerido por Hastie e Tibshirani (1990), é fixá-los de acordo com os graus efetivos de liberdade para suavização. De forma mais geral, é desejável estimar os parâmetros de suavização automaticamente, como mostraremos na Seção 2.6 (STASINOPOULOS et al., 2017).

2.5 Análise de resíduos

Em modelos de regressão linear, os resíduos ordinários são definidos pela diferença entre os valores observados e estimados, ou seja,

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}.$$

Porém, são limitados, no sentido de que são difíceis de se generalizar para outras distribuições além da normal. Para isso, Dunn e Smyth (1996) propuseram os resíduos quantílicos (aleatorizados) normalizados que são utilizados nos diagnósticos de modelos GAMLSS.

A principal vantagem dos resíduos quantílicos (aleatorizados) normalizados é que, para qualquer que seja a distribuição da variável resposta, os resíduos verdadeiros sempre terão uma distribuição normal padrão quando o modelo assumido está correto. A verificação de pressupostos através da normalidade dos resíduos já é bastante estabelecido na literatura estatística. Assim, os resíduos quantílicos (aleatorizados) normalizados nos fornecem uma maneira familiar de se verificar a adequação de um modelo (STASINOPOULOS et al., 2017). Estes resíduos são definidos por (DUNN; SMYTH, 1996)

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i),$$

em que Φ^{-1} é o inverso da distribuição acumulada de uma normal padrão e $\hat{u}_i$ são os resíduo quantílicos que são definidos de maneiras diferentes para variáveis discretas e contínuas.

Se y é uma observação de uma variável contínua então $u = F(y|\boldsymbol{\theta})$ e $\hat{u} = F(y|\hat{\boldsymbol{\theta}})$ são os valores da função na distribuição acumulada do modelo e do ajuste, respectivamente. Se o modelo está bem especificado, então u tem distribuição uniforme entre zero e um. Este processo é chamado de transformação de probabilidade integral, e é mostrado graficamente na Figura 2.2. O primeiro gráfico apresenta a distribuição de densidade de probabilidade para uma observação específica y. O segundo gráfico mostra como y é encontrada em u por meio da distribuição acumulada. No terceiro gráfico u é transformado em resíduo, chamado de *z-score*, por

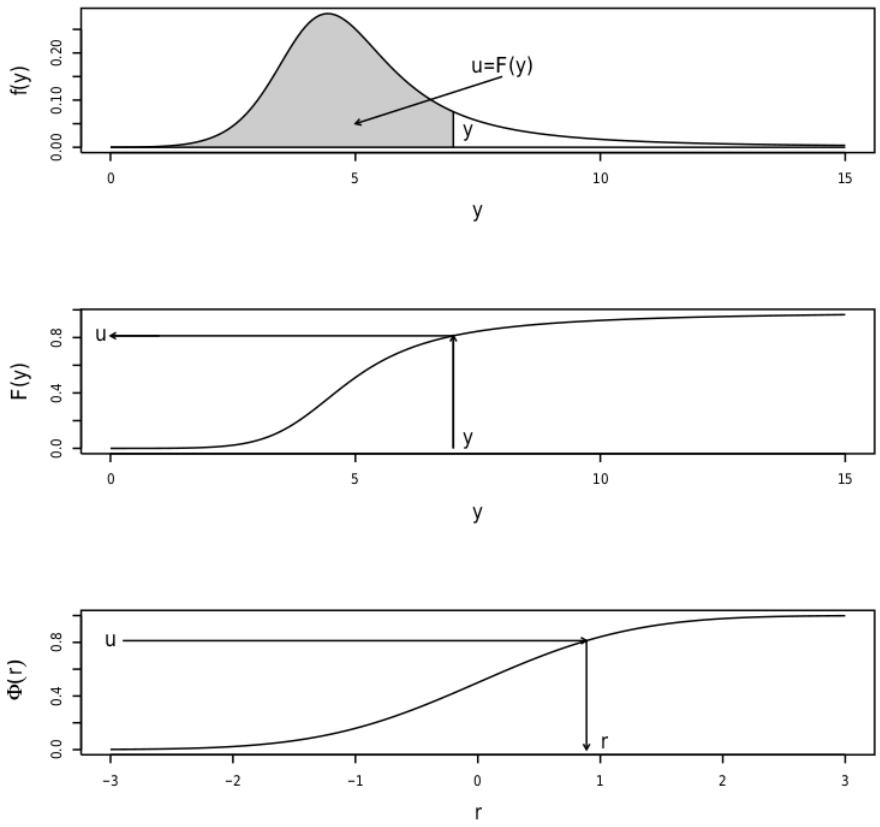
$$r = \Phi^{-1}(u),$$

que tem distribuição normal padrão se o modelo é correto. De forma similar, $\hat{u}$ é transformado nos resíduos ajustados $\hat{r}$ por

$$\hat{r} = \Phi^{-1}(\hat{u}) = \Phi^{-1}\left[F(y|\hat{\boldsymbol{\theta}})\right],$$

e $\hat{r}$ tem aproximadamente distribuição normal padrão (STASINOPOULOS et al., 2017).

Figura 2.2 – Processo de transformação de probabilidade integral contínuo (STASINOPOULOS et al., 2017, p. 420).



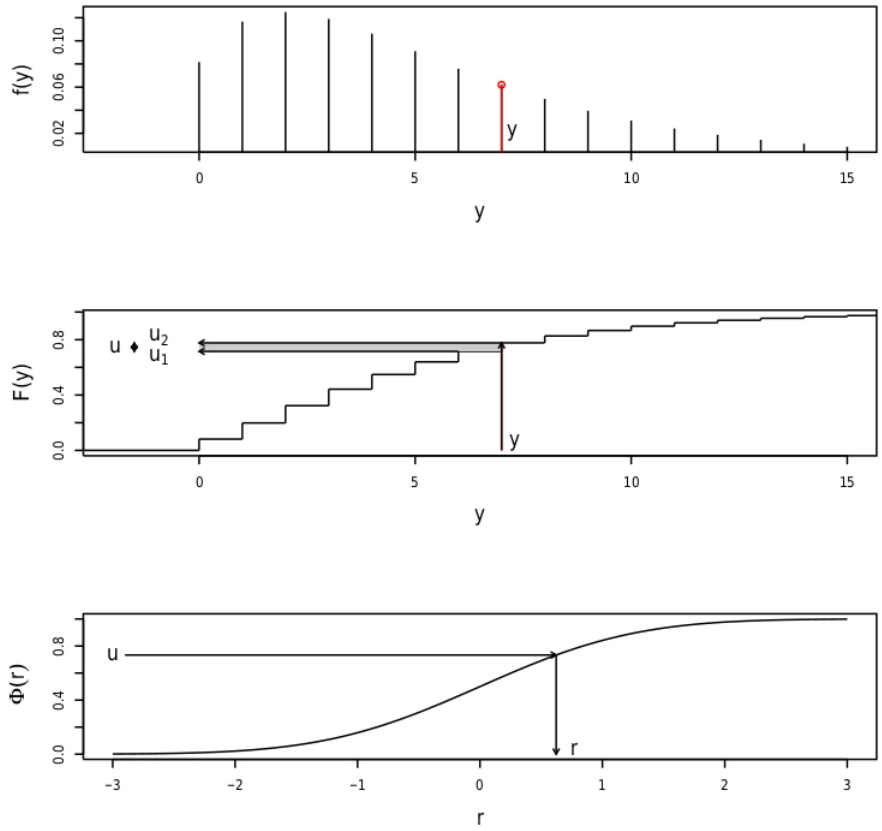
Se y é uma observação de uma variável discreta então $F(y|\boldsymbol{\theta})$ é uma função degrau. A distribuição de $u = F(y|\boldsymbol{\theta})$ tem amplitude de zero a um, mas é discreta com ponto que tem probabilidades positivas. Uma maneira de contornar esta particularidade é definir u e $\hat{u}$ como um valor aleatório de uma distribuição uniforme nos intervalos

$$\begin{matrix}[u_1, u_2] = [F(y-1|\boldsymbol{\theta}), F(y|\boldsymbol{\theta})] \\ [\hat{u}_1, \hat{u}_2] = [F(y-1|\hat{\boldsymbol{\theta}}), F(y|\hat{\boldsymbol{\theta}})]\end{matrix},$$

respectivamente. O processo é explicado na Figura 2.3. Para uma dada função de probabilidade (primeiro gráfico), o valor de y é transformado em um intervalo $[u_1, u_2]$ (segundo gráfico). Então u é selecionado aleatoriamente de uma distribuição uniforme com intervalo

$[u_1, u_2]$ e transformado em resíduo (terceiro gráfico). Se o modelo é correto então os resíduos têm distribuição normal padrão e similarmente os resíduos estimados se aproximam de uma normal padrão (STASINOPOULOS et al., 2017).

Figura 2.3 – Processo de transformação de probabilidade integral discreto (STASINOPOULOS et al., 2017, p. 421).



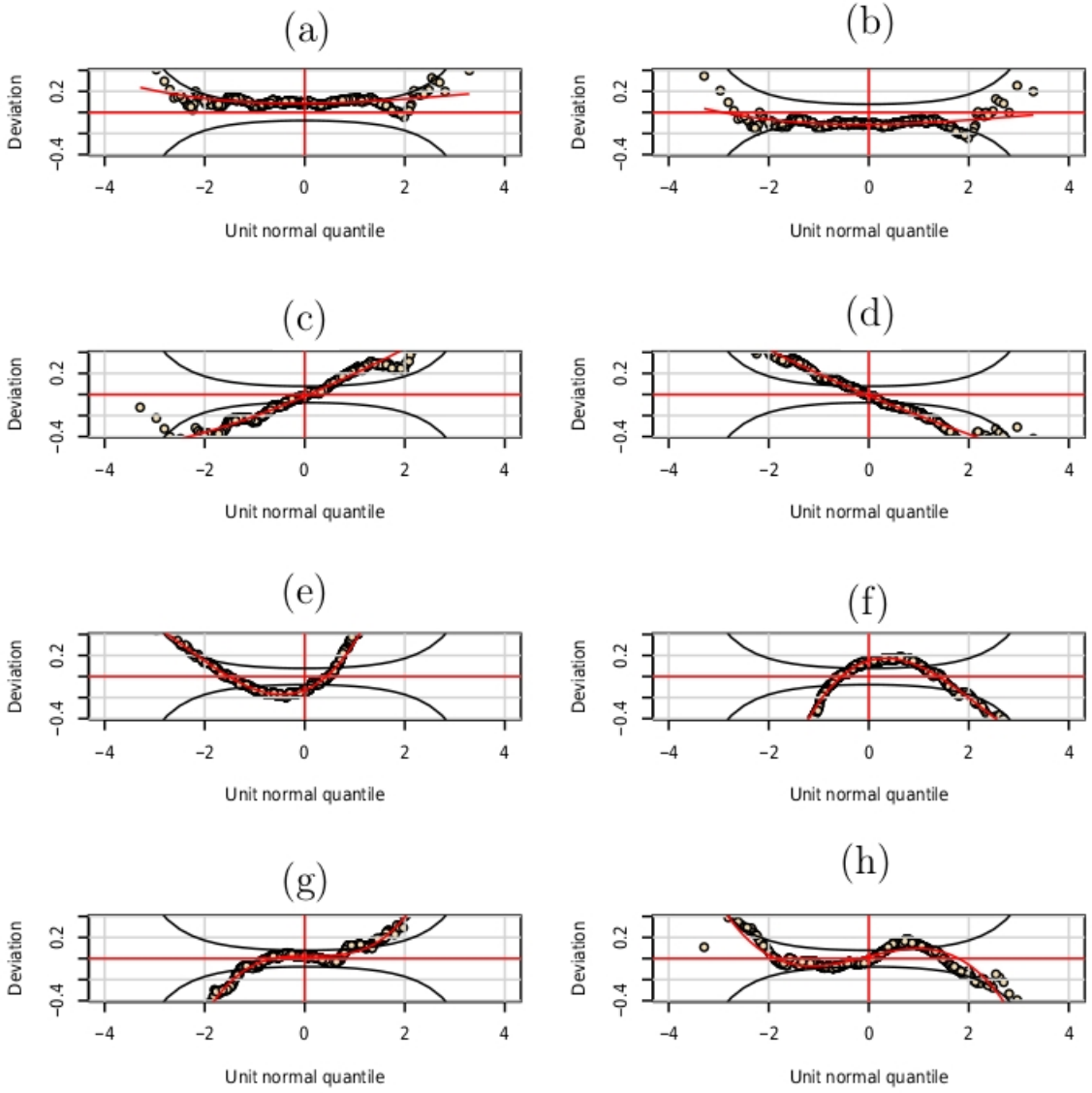
Para os casos com resposta discreta, em que a aleatorização acontece, é necessário estudar não só um conjunto de resíduos, mas vários, ou seja, várias aleatorizações, antes de tirar qualquer conclusão sobre a adequação do ajuste (STASINOPOULOS et al., 2017).

Uma das principais técnicas utilizadas para analisar os resíduos de um GAMLSS são os gráficos de minhoca (BUUREN; FREDRIKS, 2001). Esta é uma ferramenta para checar resíduos com diferentes intervalos de uma ou duas variáveis explicativas. O gráfico de minhoca pode ser considerado um gráfico normal quartil-quartil (Q-Q) sem tendência e o nome advém do formato que os pontos geralmente tem. A Tabela 2.1 mostra as possíveis interpretações para os possíveis formatos dos pontos e a Figura 2.4 mostra estes diagnósticos.

Tabela 2.1 – Diferentes formatos do gráfico de minhoca e interpretações (STASINOPOULOS et al., 2017).

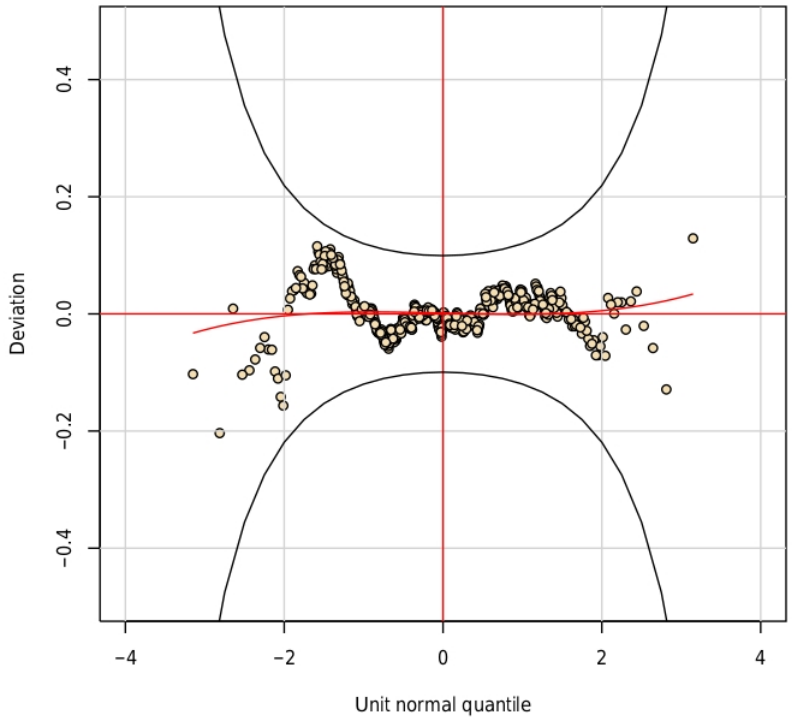
Caso	Formato	Resíduos	Parâmetro ajustado
(a)	nível: acima da origem	média muito alta	locação subestimada
(b)	nível: abaixo da origem	média muito baixa	locação superestimada
(c)	reta: inclinação positiva	variância muito alta	escala subestimada
(d)	reta: inclinação negativa	variância muito baixa	escala superestimada
(e)	U	assimetria positiva	assimetria subestimada
(f)	U invertido	assimetria negativa	assimetria superestimada
(g)	S com curva esquerda pra baixo	leptocurtose	calda muito leve
(h)	S com curva esquerda pra cima	platicurtose	calda muito pesada

Figura 2.4 – Diferentes problemas que podem ser diagnosticados pelo gráfico de minhoca (STASINOPOULOS et al., 2017, p. 429).



Se os pontos não apresentam nenhum problema citado na Tabela 2.1 e estão dentro das bandas de confiança, como na Figura 2.5, então podemos dizer que o modelo está bem ajustado.

Figura 2.5 – Gráfico de minhoca de um modelo bem ajustado (STASINOPOULOS et al., 2017, p. 427).



2.6 Análise de modelos GAMLSS via *software* R

Esta seção tem o objetivo de abordar a implementação dos modelos GAMLSS no *software* R (R Core Team, 2017). Os pacotes gamlss podem ser baixados e instalados a partir do repositório do R[1]. Atualmente, o pacote compreende o principal e uma série de outros pacotes adicionais. Stasinopoulos et al. (2017) trazem uma breve introdução sobre cada um deles:

1. O pacote original gamlss contém a principal função `gamlss()` para ajuste de modelos GAMLSS e métodos de análise de objetos resultantes dos modelos ajustados.

2. O pacote gamlss.add fornece termos aditivos extras para o ajuste dos parâmetros da distribuição da variável resposta por `gamlss()`. Este pacote também proporciona uma interface que dialoga com outros pacotes do R. Por exemplo, ao carregar gamlss.add, automaticamente pacotes como nnet, rpart e mgcv também são carregados, os quais trabalham com redes neurais, árvores de decisão e alisamentos multidimensionais, respectivamente.

[1] Disponível em: https://CRAN.R-project.org/package=gamlss

3. O pacote gamlss.cens foi criado para ajuste de variáveis respostas censuradas, seja à direita, à esquerda ou ambos, gera distribuições da `gamlss.family` que se ajustam aos modelos GAMLSS com dados censurados.

4. O pacote gamlss.data contém uma série de bancos de dados que foram utilizados em Stasinopoulos et al. (2017) e é automaticamente carregado com o gamlss.

5. O pacote gamlss.demo traz demonstrações didáticas. Ele foi criado basicamente com dois intuitos. O primeiro é mostrar visualmente como as formas das distribuições `gamlss.family` mudam a partir de diferentes valores de parâmetros. O segundo é visualizar os possíveis efeitos de suavizações.

6. O pacote gamlss.dist contém todas as distribuições disponíveis para ajuste de modelos GAMLSS e também é carregado conjuntamente com gamlss.

7. O pacote gamlss.mx permite o ajuste de distribuições de mistura e efeitos aleatórios não-paramétricos. Em Stasinopoulos et al. (2017) estes assuntos são abordados nos capítulos 7 e 10, respectivamente. Não abordaremos aqui tal particularidade.

8. O pacote gamlss.nl é utilizado para modelos GAMLSS não-lineares nos parâmetros.

9. O pacote gamlss.spatial é utilizado em modelos espaciais, quando há interesse em ajustar informações geográficas, como vizinhança ou região, via GAMLSS.

10. O pacote gamlss.tr é utilizado para truncar qualquer distribuição da `gamlss.family`, seja à direita, à esquerda ou ambos.

Os pacotes indispensáveis para análise são gamlss, gamlss.dist e gamlss.data. Os demais podem ser carregados à critério do pesquisador, de acordo com as necessidades específicas de cada ajuste. Mais informações sobre cada pacote podem ser encontradas nos arquivos de ajuda do R (`help()`). A seguir, iremos descrever as principais funções utilizadas no ajuste dos GAMLSS.

A função `gamlss()` funciona de forma análoga às conhecidas `lm()` e `glm()`. No argumento `formula` é definido o preditor de μ, em `sigma.formula` o preditor de σ, em `nu.formula` o preditor de ν e `tau.formula` o preditor de τ. No argumento `family`, definimos a distribuição desejada e podemos alterar as funções de ligação para os parâmetros pelos argumentos `mu.link=`, `sigma.link=`, `nu.link=` e `tau.link=`. O processo

iterativo que será utilizado na função `gamlss()` é definido no argumento `method`, que, se não definido pelo usuário, considera `method=RS()` por padrão. O argumento `method` também aceita mais duas entradas, `method=CG()` e `method=mixed()`, sendo que, esta última, mistura iterações dos dois algoritmos, iniciando por RS e terminando por CG. Também existem argumentos para definir algum parâmetro como constante, ou também valores iniciais para o começo das iterações. É interessante guardarmos esta função em algum objeto, para posteriormente aplicarmos outras funções e visualizar as informações do ajuste. Para obtermos estimativas para $\boldsymbol{\lambda}$ pelo método de Akaike generalizado, podemos utilizar a função `find.hyper()` (RIGBY; STASINOPOULOS, 2014).

A função `gen.likelihood()` foi criada para gerar as funções de verossimilhança do modelo com o intuito de criar a matriz hessiana utilizada na estimativa dos erros padrões. É importante destacar aqui que `gen.likelihood()` considera os termos de suavização constante, portanto, a hessiana para estes casos não leva em conta a variabilidade do ajuste das suavizações, sendo suas estimativas inválidas para este caso. Já a função `vcov()`, utiliza a função `gen.likelihood()` para obter a matriz hessiana numericamente, estimando assim a matriz de variâncias e covariâncias (STASINOPOULOS et al., 2017).

Uma outra função, muito utilizada inclusive para modelos de regressão em geral, é `summary()`, que, como o próprio nome sugere, oferece um resumo do ajuste. Esta função é mais geral porque fornece os termos do modelo, a função de ligação utilizada, os erros padrões (não levando em conta o efeito de suavização), valores-p e estatísticas t do teste de Wald para todos os termos, o tamanho amostral, os graus de liberdade do ajuste e dos resíduos, a quantidade de iterações que ocorreram até a convergência, desvio global, AIC e BIC.

Para avaliar a significância individual das variáveis explicativas, dado todos as demais no modelo, geralmente é melhor utilizar a função `drop1()` ao invés de avaliar os valores-p de `summary()`. A função `drop1()` fornece o teste de razão de verossimilhança generalizado (GLRT) para cada variável, o que é bem mais confiável que o teste de Wald baseado nos erros padrões dados pelos valores-p. Isto se aplica apenas a modelos que não possuem funções de suavização (STASINOPOULOS et al., 2017).

A função `confint()` também já é bastante utilizada em modelos lineares e fornece intervalos de confiança baseados em erros padrões de Wald para os coeficientes ajustados

aos parâmetros da distribuição, seja μ, σ, ν e τ. No pacote gamlss, foram implementadas duas funções para obter intervalos de confiança: `prof.dev()` e `prof.term()`.

A primeira, `prof.dev()`, fornece um gráfico de perfil de desvio, útil para avaliar a confiabilidade de modelos em que um ou mais parâmetros da distribuição sejam constantes, ou seja, não modelados por variáveis explicativas. Além disso, `prof.dev()` também proporciona um intervalo de confiança baseado em verossimilhança perfilada para o parâmetro constante, que em geral é bem mais confiável que `confint()` se temos um modelo bem ajustado. Já a função `prof.term()` é semelhante à `prof.dev()`, porém, ela fornece o perfil de desvio para qualquer parâmetro da distribuição, inclusive se ele for ajustado em função de variáveis explicativas É importante destacar que intervalos de confiança por perfil de desvio para coeficientes com efeitos aleatórios ou funções de suavização devem ser estimados com cuidado.

Em geral, para estes casos, a função `prof.term()` pode produzir intervalos muito estreitos, por não levar em conta esta variabilidade extra. Intervalos mais precisos são obtidos por aproximações da verossimilhança marginal, que depende do modelo ajustado. Até o momento não existe função implementada no R para ajustar intervalos de confiança neste caso (STASINOPOULOS et al., 2017).

A função `predict()`, que em gamlss é baseada na função `predict.gam()` que utiliza linguagem S-PLUS, produz predições de dados para algum parâmetro específico do modelo. Stasinopoulos et al. (2017) também implementaram a função `predictAll()` que produz previsões para todos os parâmetros da distribuição do modelo `gamlss()`. O usuário pode precisar alterar os valores iniciais para os parâmetros da distribuição μ, σ, ν e τ, usando os argumentos `mu.start`, `sigma.start`, `nu.start` e `tau.start`, respectivamente.

Com relação às funções de suavização que podemos aplicar às variáveis explicativas, todas que estão implementas hoje no pacote gamlss são elencadas na Tabela 2.2, com as suas respectivas funções em R, que são aplicadas dentro do argumento `formula` dos parâmetros.

Existem muitas estratégias diferentes que poderiam ser aplicadas para a seleção de variáveis explicativas usadas para modelar todos os parâmetros $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$. Na implementação do pacote gamlss, atualmente, podemos destacar duas estratégias para selecionar termos para todos os parâmetros, que Stasinopoulos et al. (2017) chamaram de estratégia A e estratégia B. Elas estão implementadas nas funções `stepGAICAll.A()` e

Tabela 2.2 – Termos aditivos implementados no pacote gamlss (STASINOPOULOS et al., 2017).

Termos Aditivos	Função em gamlss
splines cúbicas	`cs()`, `scs()`
árvore de decisão	`tr()`
polinômios fracionais e potência	`fp()`, `pp()`
free knots (break points)	`fk()`
`loess`	`lo()`
redes neurais	`nn()`
ajuste não linear	`nl()`
P-splines	`pb()`, `pb0()`, `ps()`
P-splines cíclicos	`pbc()`, `cy()`
P-splines monótonos	`pbm()`
P-splines encolhidos em zero	`pbz()`
P-splines de coeficientes variantes	`pvc()`
categórico penalizado	`pcat()`
efeitos aleatórios	`re()`, `random()`
regressão *ridge*	`ri()`

`stepGAICAll.B()`, respectivamente. Por padrão, estas funções nos fornecem o melhor modelo para um dada distribuição segundo o AIC, mas o critério pode ser alterado nos argumentos da função.

A função mais utilizada é a `stepGAICAll.A()` e ela acontece da seguinte maneira. Primeiro, é ajustado um modelo para μ por meio do procedimento *forward* para algum GAIC selecionado, considerando os demais parâmetros constantes. Em seguida, é ajustado para σ, pelo mesmo procedimento, considerando o primeiro modelo para μ e demais constantes. Se houverem mais parâmetros, então ν é ajustado, considerando os modelos de μ e σ. Este procedimento ocorre até que o último parâmetro da distribuição seja modelado. A seleção de modelos para o último parâmetro da distribuição ocorre uma única vez, de forma que, após todos os parâmetros ajustados por *forward*, o algoritmo começa a retroceder, reajustando o penúltimo parâmetro, pelo procedimento *backward*, dado todos os demais já ajustados, até retornar no reajuste de μ (NAKAMURA et al., 2017). Ao fim deste processo, o algoritmo pára e os modelos para cada parâmetro compõem o modelo GAMLSS final, considerado o mais adequado pelo critério escolhido. Salientamos aqui, ainda, que diferentes covariáveis podem aparecer em cada estrutura de regressão.

Como já descrito anteriormente, os resíduos quantílicos (aleatorizados) normalizados são utilizados na análise de resíduos. Estes podem ser obtidos no pacote gamlss por meio da função `resid()`. Existem outras funções que auxiliam no diagnóstico a partir de tais resíduos. A função `plot()`, amplamente utilizada em análises de modelos lineares, em

GAMLSS produz quatro gráficos para avaliarmos os resíduos quantílicos (aleatorizados) normalizados de um objeto `gamlss` (STASINOPOULOS et al., 2017): resíduos *versus* valores ajustados do parâmetro μ, resíduos *versus* indexação da ordem das observações, uma estimação da densidade kernel dos resíduos e um gráfico normal Q-Q. A função `wp()` gera os gráficos de minhoca que foram mencionados na Seção 2.5 e a função `rqres.plot()` é usada para criar múltiplas realizações dos resíduos quantílicos (aleatorizados) normalizados quando a distribuição da variável resposta é discreta, e esboçá-los usando gráficos de minhoca ou Q-Q.

Mais detalhes sobre o ajuste de modelos GAMLSS em R podem ser encontrados em Stasinopoulos et al. (2017).

2.7 Aplicações pré-existentes dos GAMLSS na experimentação agropecuária

Nesta seção, serão descritos, resumidamente, alguns trabalhos relacionados ao uso dos modelos GAMLSS aplicados a dados de experimentação agropecuária.

No trabalho de Righetto et al. (2018), os autores utilizaram um modelo de regressão multinomial, por meio das técnicas GAMLSS, para predizer a infestação de ervas daninhas em uma plantação de cana-de-açúcar por meio de imagens multiespectrais. A distribuição multinomial é uma generalização da distribuição binomial e pode ser utilizada quando a variável resposta é discreta e nominal e assume mais de duas classificações possíveis. Os modelos multinomiais são utilizados com o objetivo de estimar a probabilidade da ocorrência de cada classificação em função de variáveis explicativas. Com a utilização deste modelo, os autores puderam predizer dados faltantes, de forma que, em toda plantação, foi estimado que o uso de herbicidas poderia ser reduzido de 84% para 27%, que é uma vantagem econômica na medida que é possível a redução de custos com esses químicos, e também um ganho social, na medida que a produção pode se desenvolver de forma mais natural. No artigo, os autores disponibilizam a rotina utilizada em R (R Core Team, 2017), de forma a facilitar ajustes semelhantes para outras cultivares, por exemplo.

Eloy (2016), em seu trabalho, propôs a utilização de um modelo de regressão com distribuição *t-Student* assimétrica tipo 3 para avaliar o efeito do teor de boro e a absorção de enxofre na produção de grãos de soja, considerando dados de experimentos instalados segundo o delineamento inteiramente casualizado. Essa distribuição possui quatro parâmetros e é utilizada para dados contínuos. Ela pode ter formatos assimétricos

ou não, platicúrticos ou leptocúrticos, unimodal ou bimodal, a depender dos valores dos parâmetros. A autora pôde concluir que tais modelos podem ser utilizados na análise de dados obtidos por delineamentos experimentais com interpretações práticas, sem que haja a necessidade de pressupostos muito rígidos, como ocorre nas análises de variância. Também, esta distribuição se adequou melhor aos dados em relação à distribuição normal.

Peixoto e Sage (2016) verificaram a tolerância ao frio de dois tipos de grama por meio de regressão beta, binomial e modelos lineares generalizados mistos. A condutividade relativa, uma das variáveis respostas estudadas, apresenta valores entre 0 e 1, de forma que é adequado ajustar um modelo beta, que, neste trabalho, foi feito via GAMLSS. Os demais modelos também poderiam ser estimados por GAMLSS, entretanto, os autores utilizam apenas a teoria de modelos mistos. Nesse trabalho não havia o interesse em detalhar as análises estatísticas, porém, podemos destacar que os GAMLSS estão sendo utilizados não só por estatísticos, mas também por pesquisadores de diversas áreas.

Barajas et al. (2015) ajustaram modelos de regressão gama, modelando a média e variância em função de covariáveis, por meio de modelos GAMLSS, para estudar a produção de celulose bacteriana a partir de rejeitos agroindustriais. O experimento foi realizado com intuito de investigar os efeitos do pH e o tempo de cultivo sobre o rendimento de celulose bacteriana obtida por meio de resíduos da banana. Eles puderam concluir que a relação entre média e a variância do rendimento da celulose e o pH é inversa, uma vez que, à medida que o pH aumenta, o rendimento da celulose diminui. O contrário ocorre para o tempo de cultivo, ou seja, à medida que aumenta o tempo de cultivo, também aumenta-se o rendimento da celulose.

O trabalho de Abbas et al. (2013), também não têm grande ênfase estatística, entretanto, os autores exploraram por meio da modelagem simultânea da média e variância, considerando distribuição normal, a estequiometria em tecidos vegetais acima do solo, como razões de carbono, nitrogênio, fósforo e potássio, e sua relação com vários tipos de plantas. Os autores comparam os resultados obtidos por GAMLSS e análise de variância (ANOVA) e concluíram que os resultados são bastante divergentes. Eles concluíram que a ANOVA fornece o teste mais poderoso na análise de contrastes, enquanto que a análise via GAMLSS fornece testes mais poderosos para alterações simultâneas na variância.

Piekarska-Boniecka et al. (2010) construíram modelos para descrever a expectativa de vida de fêmeas de moscas-serra, incluindo a análise da contagem de ovos, considerando

distribuições Poisson, binomial negativa e Poisson-normal inversa. No trabalho, as autoras consideraram a expectativa de vida como uma variável discreta de contagem, contendo os tamanhos de vida dos insetos. O melhor modelo ajustado foi o modelo que considerou à distribuição Poisson-normal inversa.

Por último, temos o trabalho de Rocha et al. (2009), cujos autores usaram modelos com distribuições beta inflacionadas, gama, normal inversa zero ajustada para determinar os efeitos de alguns fungos no cultivo de milho em diferentes regiões do Brasil. Os modelos GAMLSS foram utilizados para determinar o efeito de diferentes regiões estudadas e estimar o crescimento dos fungos. Os autores destacaram que estes estudos são necessários para determinar regiões em que são mais ou menos afetadas pelos fungos, com o objetivo de reduzir riscos de contaminação humana e animal.

3 MATERIAIS E MÉTODOS

3.1 Experimento analisado

Neste trabalho iremos analisar dados provenientes de um experimento executado por Blanco (2015). De acordo com a autora, as doenças parasitárias em bovinos ainda são um grande problema para o mercado pecuário, uma vez que os animais doentes fornecem carnes e leites contaminados impróprios para venda e consumo, além da doença ocasionar possível morte precoce do animal e contaminação de todo um rebanho ou dos trabalhadores que atuam no manejo, o que causa perdas significativas para o produtor. Além disso, a autora ressalta que "o uso indiscriminado desses produtos (antiparasitários), devido ao fácil acesso pelo produtor e a sua utilização sem orientação técnica adequada, têm aumentado o aparecimento de resistência nos parasitas"(BLANCO, 2015, p. 26).

Na busca de amenizar este problema, Blanco (2015) realizou um experimento no qual propunha avaliar a eficácia e custos de um tratamento estratégico seletivo em comparação ao tratamento convencional no controle das principais parasitoses gastrointestinais em bezerras leiteiras. O tratamento estratégico seletivo (TE) e o tratamento convencional (TC) serão descritos na seção 3.1.

A eficácia de cada tratamento pode ser medida por meio da redução da contagem de ovos por grama de fezes (OPG), que será a variável resposta, escolhida pela difícil modelagem devido ao excesso de zeros e à grande variabilidade das contagens observadas, com valores bem discrepantes.

Em seu trabalho, Blanco (2015) realizou uma análise transformando a variável original contagem de OPG em uma variável binária, considerando contagens abaixo de 300 como negativo para a doença e acima como positivas. Foi uma análise válida, mas realizar tal transformação nos dados gera uma grande perda de informações. Neste trabalho, modelamos a contagem OPG sem nenhuma transformação e, também, analisamos o possível efeito dos tratamentos. Os dados coletados neste experimento foram gentilmente fornecidos pela autora.

O experimento foi realizado em uma fazenda experimental da Universidade Federal de Lavras (UFLA) no período de abril de 2013 a novembro de 2014. Foram utilizadas 30 bezerras da raça holandesa que foram divididas aleatoriamente em dois grupos, o pri-

meiro para receber o tratamento estratégico seletivo (TE) e o segundo para o tratamento convencional (TC).

O TE foi designado pelos pesquisadores que definiram quais os medicamentos que seriam utilizados no tratamento e a frequência. A particularidade principal deste tratamento é que as bezerras eram medicadas apenas se a contagem de OPG fosse igual ou superior à 300. Já os animais que foram designados a receber o TC foram medicados com produtos da própria fazenda experimental e ministrados pelos responsáveis da propriedade.

Informações sobre as bezerras e amostras de fezes foram coletadas a cada 15 dias, o que originou uma amostra com 680 observações das quais 14 observações foram eliminadas por terem dados faltantes, totalizando 666 observações. As variáveis explicativas candidatas a serem incluídas nos preditores são:

Tratamentos: fator que indica qual tratamento a bezerra está recebendo, TE ou TC;

Faixa etária: fator que designa em qual faixa etária a bezerra estava na data da coleta, distribuída em três níveis: 0-90 dias, 91-180 dias e 181-365 dias;

Estação do ano: fator que designa em qual estação do ano foi feita a coleta: verão, outono, inverno ou primavera;

Peso: variável quantitativa que mostra o peso da bezerra no momento da coleta;

Altura: variável quantitativa que mostra a altura do animal no momento da coleta;

Temperatura: variável quantitativa que fornece a temperatura corporal da bezerra no momento da coleta;

3.2 Análise estatística

A contagem OPG foi analisada por meio das técnicas de modelagem GAMLSS. Primeiramente, foram ajustadas marginalmente todas as distribuições de contagem disponíveis no pacote gamlss pela função `fitDist()`. Em seguida, selecionamos seis distribuições de contagem diferentes: Poisson (PO), binomial negativa (NBI), Poisson-normal inversa (PIG), Poisson com inflação de zeros (ZIP), binomial negativa com inflação de

zeros (ZINBI) e Poisson-normal inversa com inflação de zeros (ZIPIG). A justificativa desta escolha é apresentada na Seção 4.1.

A função de probabilidade de uma distribuição Poisson, pode ser descrita como (JOHNSON; KEMP; KOTZ, 2005; RIGBY et al., 2017)

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!},$$

em que $y = 0, 1, 2, ...$ e $\mu > 0$. O parâmetro μ também é chamado de parâmetro de intensidade e corresponde à média e também à variância da distribuição, sendo esta a principal característica da distribuição Poisson uniparamétrica.

Na literatura, podemos encontrar algumas parametrizações para a função de probabilidade da distribuição binomial negativa, uma delas é (JOHNSON; KEMP; KOTZ, 2005; RIGBY et al., 2017)

$$P(Y = y) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(y + 1)\Gamma(\frac{1}{\sigma})} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^{\frac{1}{\sigma}} \tag{3.1}$$

em que $0 < \mu < \infty$ é a média da distribuição e $0 < \sigma < \infty$ é o parâmetro de dispersão. A variância de uma distribuição binomial negativa nesta parametrização é $\mu + \sigma\mu^2$. A principal característica desta distribuição é que ela é capaz de modelar superdispersão, quando a variância da amostra é superior à média.

A distribuição Poisson-normal inversa também possui diferentes parametrizações. A parametrização adotada por Dean, Lawless e Willmot (1989), Rigby et al. (2017) é

$$P(Y = y) = \left(\frac{2\alpha}{\pi}\right)^{1/2} \frac{\mu^y e^{1/\sigma} K_{y-\frac{1}{2}}(\alpha)}{y!(\alpha\sigma)^y}, \tag{3.2}$$

sendo $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$ e $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\left\{-\frac{1}{2}t(x + x^{-1})\right\} dx$, em que K é a função Bessel do terceiro tipo modificada, $\mu > 0$ é a média da distribuição e $\sigma > 0$ é o parâmetro de dispersão. A variância também é dada por $\mu + \sigma\mu^2$. Esta distribuição é uma boa alternativa, em detrimento da binomial negativa, quando a distribuição possui caudas muito longas (DEAN; LAWLESS; WILLMOT, 1989).

Estas três distribuições apresentadas podem ter sua variação para adequar a situações em que ocorrem excesso de zeros. As distribuições com inflação em zeros são, em sua essência, a combinação de duas distribuições conhecidas. Por este motivo, estes modelos

também são conhecidos por modelos de mistura. Assim, ajusta-se a contagem, que pode ser feita por distribuição de Poisson, binomial negativa ou outra e ajusta-se também uma probabilidade da ocorrência de zero pela distribuição Bernoulli dentro do mesmo modelo. Assim, há a contribuição para a estimação da probabilidade em zero de duas funções de probabilidade (ZEVIANI; JR; TACONELI, 2016).

Uma distribuição Poisson com inflação de zeros (ZIP) pode ser dada por (LAMBERT, 1992) e (RIGBY et al., 2017)

$$P(Y=y)=\begin{cases}\sigma+(1-\sigma)e^{-\mu}, & se \;\; y=0\\ (1-\sigma)\frac{e^{-\mu}\mu^{y}}{y!}, & se \;\; y=1,2,...\end{cases}$$

em que $0<\mu<\infty$ é a média da componente Poisson e $0<\sigma<1$ é a probabilidade da ocorrência de zeros ($Y=0$). A média dessa distribuição é $(1-\sigma)\mu$ e variância $\mu(1-\sigma)(1+\mu\sigma)$.

Para as distribuições binomial negativa e Poisson-normal inversa, a inflação de zeros é dada por (RIGBY et al., 2017)

$$P(Y=y)=\begin{cases}\nu+(1-\nu)P(Y_1=0), & se \;\; y=0\\ (1-\nu)P(Y_1=y), & se \;\; y=1,2,...\end{cases}$$

em que, para ambas distribuições, $0<\mu<\infty$ e $0<\sigma<\infty$ são à média e a dispersão da componente da distribuição de contagem e $0<\nu<1$ é a probabilidade extra de $Y=0$. Aqui, $Y_1\sim NB(\mu,\sigma)$ ou $Y_1\sim PIG(\mu,\sigma)$, como definido, respectivamente, nas Equações 3.1 e 3.2, então $Y\sim ZINBI(\mu,\sigma,\nu)$ ou $Y\sim ZIPIG(\mu,\sigma,\nu)$. Para ambos os casos, a média da distribuição é dada por $(1-\nu)\mu$ e variância $\mu(1-\nu)+\mu^2(1-\nu)(\sigma+\nu)$.

Na análise, esperamos que a distribuição ZIPIG apresente melhores ajustes, uma vez que a natureza da variável da contagem OPG amostrada apresenta excesso de zeros e caudas longas.

3.3 *Software*

Todas as análises, resultados e gráficos apresentados no presente trabalho foram obtidos por meio da linguagem de programação R (R Core Team, 2017). A linguagem R é gratuita e amplamente utilizada para análises estatísticas no meio acadêmico e toda a classe de modelos GAMLSS está implementada neste ambiente por meio de uma série

de pacotes e seus respectivos manuais, que estão disponíveis para download também gratuitamente no próprio repositório do R. Foram utilizados os pacotes `gamlss` (STASINOPOULOS; RIGBY, 2007) e `moments` (KOMSTA; NOVOMESTKY, 2015). No Apêndice A deste trabalho, é mostrado a rotina em R utilizada para as devidas consultas de interesse.

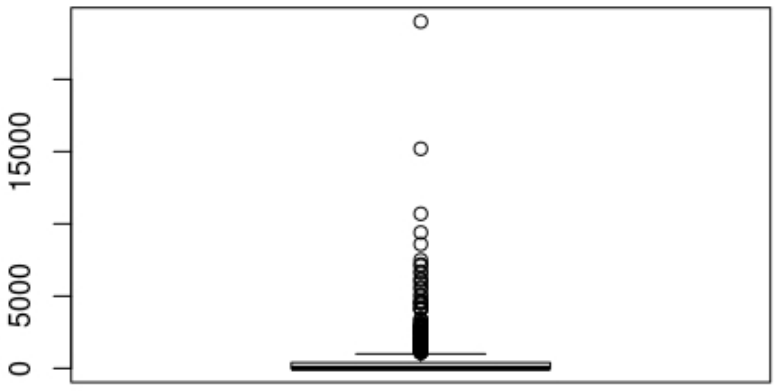
4 RESULTADOS E DISCUSSÃO

Neste capítulo, apresentamos uma análise exploratória dos dados, iniciando pela a variável resposta, em seguida, as variáveis explicativas quantitativas e fatores. Em seguida, apresentamos os resultados do ajuste da contagem OPG, mostrando os modelos construídos e algumas inferências. Por fim, também avaliamos o efeito dos tratamentos no modelo final.

4.1 Análise exploratória

A variável resposta, contagem OPG, como já mencionada na Seção 3.1, é uma variável discreta, que tem valores observados no intervalo dos inteiros positivos. O mínimo desta variável é 0 e o máximo 24.000. Na Figura 4.1, temos um gráfico de caixa da contagem OPG, com muitos valores nulos e algumas observações elevadas. A média para esta variável é $605,33$ e o desvio padrão é $1670,30$. Assim já se tem um forte indício de que a distribuição Poisson não se adequará bem, uma vez que ela pressupõe igualdade de médias e variâncias. O primeiro quartil e mediana são nulos, o que leva à informação de que mais da metade das observações são nulas. O terceiro quartil é igual à 400, ou seja, 75% das observações estão abaixo deste valor. Podemos observar que a distribuição é assimétrica, apresentando coeficiente de assimetria positivo igual à $6,82$. Além disso, o coeficiente de curtose foi igual à $73,77$, indicando que a distribuição destes dados é fortemente leptocúrtica. O elevado número de observações atípicas são características da variável, ou seja, não são erros de mensuração.

Figura 4.1 – Gráfico de caixas para a contagem de OPG.



Por meio da função `fitDist()`, utilizada para ajustar distribuições marginais para a variável resposta, foram ajustados marginalmente todas as distribuições discretas de contagem implementadas no pacote `gamlss`. Assim, podemos ter indícios de quais serão as melhores e piores distribuições de probabilidade condicionais para OPG. O resultado encontra-se na Tabela 4.1 em ordem crescente dos valores AIC, de forma que podemos observar uma predominância de melhores ajustes para distribuições que englobam o excesso de zeros.

A distribuição Poisson (PO) foi a que obteve maior AIC e a Poisson-normal inversa com inflação de zeros (ZIPIG) foi a que obteve o menor. As distribuições foram denotadas pelas siglas utilizadas no pacote `gamlss` e podem ser consultadas no Apêndice B.

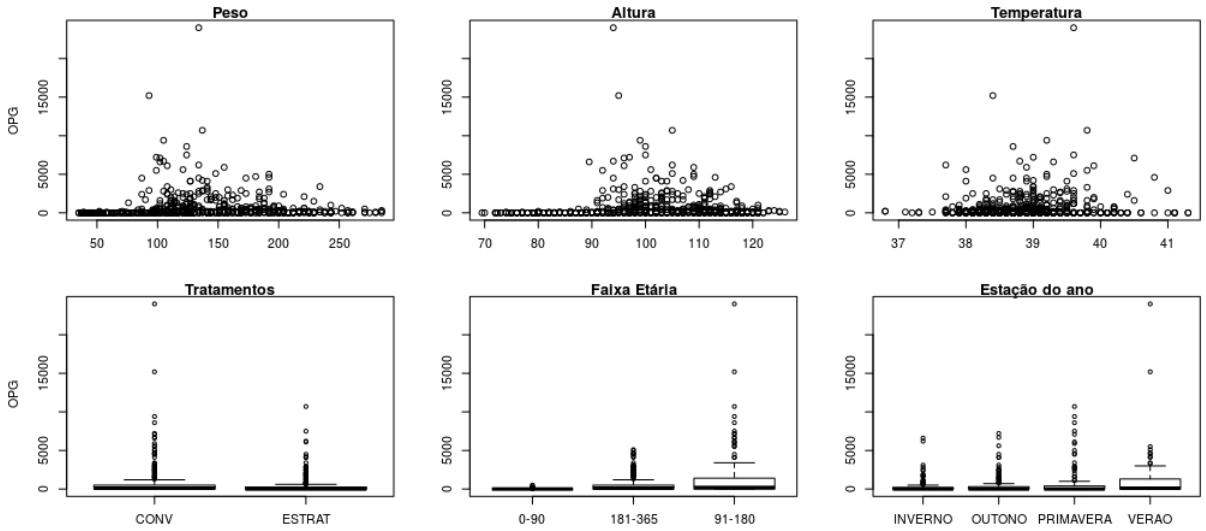
Tabela 4.1 – Distribuições marginais de contagem para contagem OPG e seus respectivos AIC.

Distribuição	AIC	Distribuição	AIC
ZAPIG	6017,04	SI	6372,71
ZIPIG	**6017,04**	ZALG	6467,02
ZASICHEL	6019,00	ZAZIPF	6835,66
ZISICHEL	6019,00	GPO	7317,69
ZABNB	6042,38	**PIG**	**7538,60**
ZIBNB	6042,38	DPO	7573,26
ZINBI	**6146,95**	WARING	8244,93
ZANBI	6146,95	YULE	8526,96
ZINBF	6148,95	GEOM	9867,59
NBI	**6370,65**	GEOMo	9867,59
NBII	6370,65	**ZIP**	**686561,91**
NBF	6372,65	ZIP2	686561,91
SICHEL	6372,65	ZAP	686561,91
DEL	6372,65	**PO**	**1266613,49**
BNB	6372,65		

Em GAMLSS, temos muitas opções para a distribuição da variável resposta, sendo que `fitDist()` nos orienta sobre quais escolher à depender do interesse do pesquisador. Neste trabalho, como já mencionado na Seção 3.2, escolhemos 6 distribuições para modelar a contagem OPG: Poisson, binomial negativa, Poisson-normal inversa e suas variações para inflação de zeros. Observando a Tabela 4.1, podemos notar que, para a contagem OPG, selecionamos distribuições adequadas, inadequadas e intermediárias. Isto foi feito para realizarmos comparações e descrever detalhadamente como são realizadas as análises e como são interpretados os resultados ao utilizar GAMLSS. Além disso, teve-se a preocupação de escolher distribuições que nos forneçam interpretações práticas e aplicadas para os parâmetros.

Na Figura 4.2 podemos ver o comportamento das contagens OPG para as variáveis explicativas por meio dos gráficos de dispersão. Com relaçao ao peso, percebemos que as contagens mais altas encontram-se na faixa de 100kg à 150kg. Com relação à altura, observamos contagens elevadas entre 90cm e 110cm. Já em relação à temperatura, percebemos os pontos mais altos no intervalo de 38^oC e 41^oC. Os fatores que apresentam maiores contagens são: tratamento convencional, faixa etária de 91-180 dias e estação do ano verão. Em todos os gráfico podemos perceber o comportamento do excesso de zeros.

Figura 4.2 – Gráfico de dispersão da contagem de OPG e variáveis explicativas.



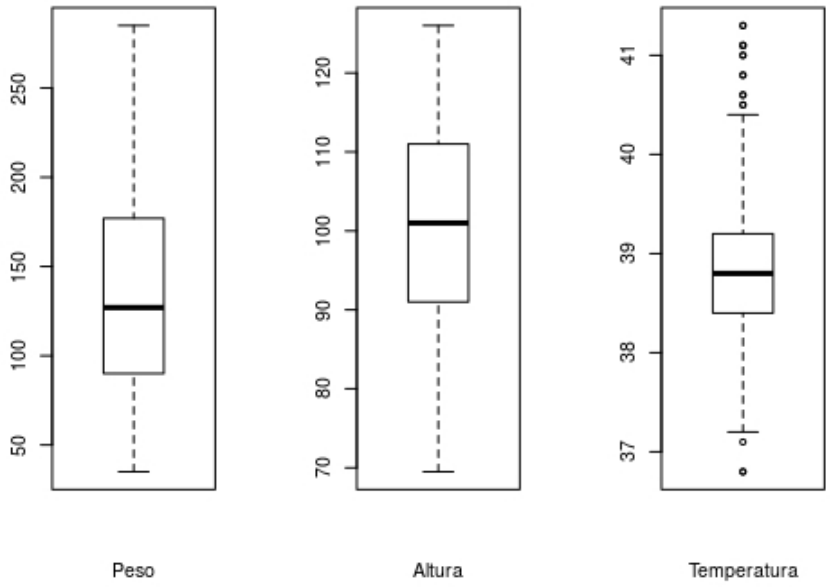
Na Tabela 4.2 podemos visualizar algumas estatísticas descritivas das variáveis explicativas quantitativas: altura, peso e temperatura. Figura 4.3 podemos ver os gráficos de caixa de tais variáveis. A variável peso possui variância bastante alta se comparada com a altura. Peso e temperatura apresentam leve assimetria positiva enquanto que altura possui leve assimetria negativa. As variáveis não apresentam grandes problemas de curtose, uma vez que os valores são próximos à 3, exceto pela temperatura, que é levemente leptocúrtica, porque possui 8 observações discrepantes, conforme Figura 4.3.

É sabido que as variáveis altura e peso são altamente correlacionadas. Para este caso o coeficiente de correlação de Pearson (RODGERS; NICEWANDER, 1988) é igual a $0,95$. Na Figura 4.4, podemos observar um gráfico de dispersão que confirma a alta correlação, uma vez que podemos observar que os pontos distribuem-se quase em uma diagonal. Para evitar o problema de colinearidade das variáveis no modelo, daremos preferência para a variável peso, porque é uma medida mais comumente utilizada na área de pesquisa com bovinos.

Tabela 4.2 – Estatísticas descritivas das variáveis quantitativas.

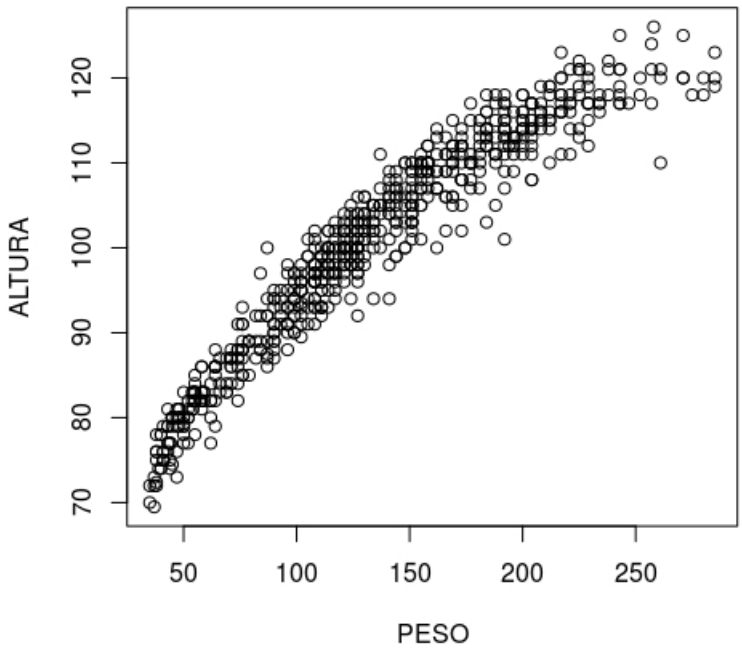
	Peso	Altura	Temperatura
Mínimo	35,00	69,50	36,80
1º Quartil	90,75	91,00	36,40
Mediana	127,00	101,00	38,80
3º Quartil	177,00	111,00	39,20
Máximo	285,00	126,00	41,30
Média	133,58	100,3	38,87
Variância	3353,47	162,36	0,41
Desvio Padrão	57,91	12,74	0,64
Coeficiente de assimetria	0,29	-0,30	0,61
Coeficiente de curtose	2,38	2,19	4,36

Figura 4.3 – Gráfico de caixas para as variáveis quantitativas.



A seguir, apresentaremos informações sobre os fatores: tratamentos, faixa etária e estação do ano. Existem 306 observações referentes ao tratamento convencional, o que representa 45,95% da amostra, sendo 360 observações referentes ao tratamento estratégico, representando 54,05% da amostra. 46,39% da amostra foi coletada de bezerras com mais de 180 dias de idade (309 observações), 26,27% de bezerras com menos de 90 dias de idade (175 observações) e 27,32% de bezerras entre estas duas idades, em um total de 182 observações. A variável estação também não divide-se igualmente. Foram coletadas 165 observações no inverno (24,77%), 185 no outono (27,78%), 172 na primavera (25,82%) e 144 no verão (21,62%).

Figura 4.4 – Peso *versus* Altura.



4.2 Ajuste e comparação dos modelos GAMLSS

O intuito desta seção é mostrar o ajuste dos modelos GAMLSS para a contagem OPG, considerando, para a seleção de variáveis, todas as cinco variáveis explicativas remanescentes: tratamento, faixa etária, estação, peso e temperatura. Os modelos apresentados a seguir foram selecionados pela função `stepGAICAll.A()`, utilizando o critério AIC e `method=mixed()`. O máximo de iterações estipuladas para os algoritmos RS e CG podem ser consultadas no Apêndice A. Para as variáveis explicativas contínuas, aplicamos a função de suavização *P-Spline*. O principal interesse dessa seção é comparar os ajustes obtidos.

A distribuição Poisson e sua variação para inflação de zeros se mostraram completamente inadequados para este banco de dados, porque a função `stepGAICAll.A()` ajustou estimativas incoerentes para os parâmetros. Observando a Tabela 4.3, ambas apresentaram valores de desvio global, AIC e BIC elevados e muito discrepantes das demais distribuições. Este era um resultado esperado, uma vez que as próprias distribuições marginais (Tabela 4.1) se ajustaram pobremente aos dados.

Sendo assim, consideramos apenas os modelos NBI, PIG, ZINBI e ZIPIG. A Tabela 4.3 mostra que os modelos inflacionados em zero (ZIPIG e ZINBI) se adequam melhor aos dados, porque apresentaram os menores valores para todos os critérios. É interessante

notar que, para as distribuições adequadas, a ordem se manteve a mesma das distribuições marginais (Tabela 4.1), mostrando como é útil utilizar a função `fitDist()` com o objetivo de orientar na seleção da distribuição.

Tabela 4.3 – Critérios de seleção para os modelos da contagem OPG.

Distribuição	Desvio global	AIC	BIC
ZIPIG	5711,44	5743,45	5815,49
ZINBI	5772,06	5823,97	5940,81
NBI	6031,61	6079,91	6188,61
PIG	7106,96	7145,56	7232,42
PO	696983,00	697025,90	697122,20
ZIP	3883117,00	3883220,00	3883454,00

Na Tabela 4.4, elencamos quais variáveis explicativas foram inclusas nos preditores das distribuições. Não é interessante aqui comparar as estimativas dos coeficientes porque a maneira como as variáveis influenciam nos parâmetros muda de acordo com as variáveis explicativas presentes. A principal observação a ser feita é que a variável tratamento não foi significativa no modelo ZINBI. Todos os demais modelos tiveram todas as covariáveis inclusas, considerando todos os preditores.

Tabela 4.4 – Variáveis inclusas nos preditores dos parâmetros dos modelos por `stepGAICAll.A()`.

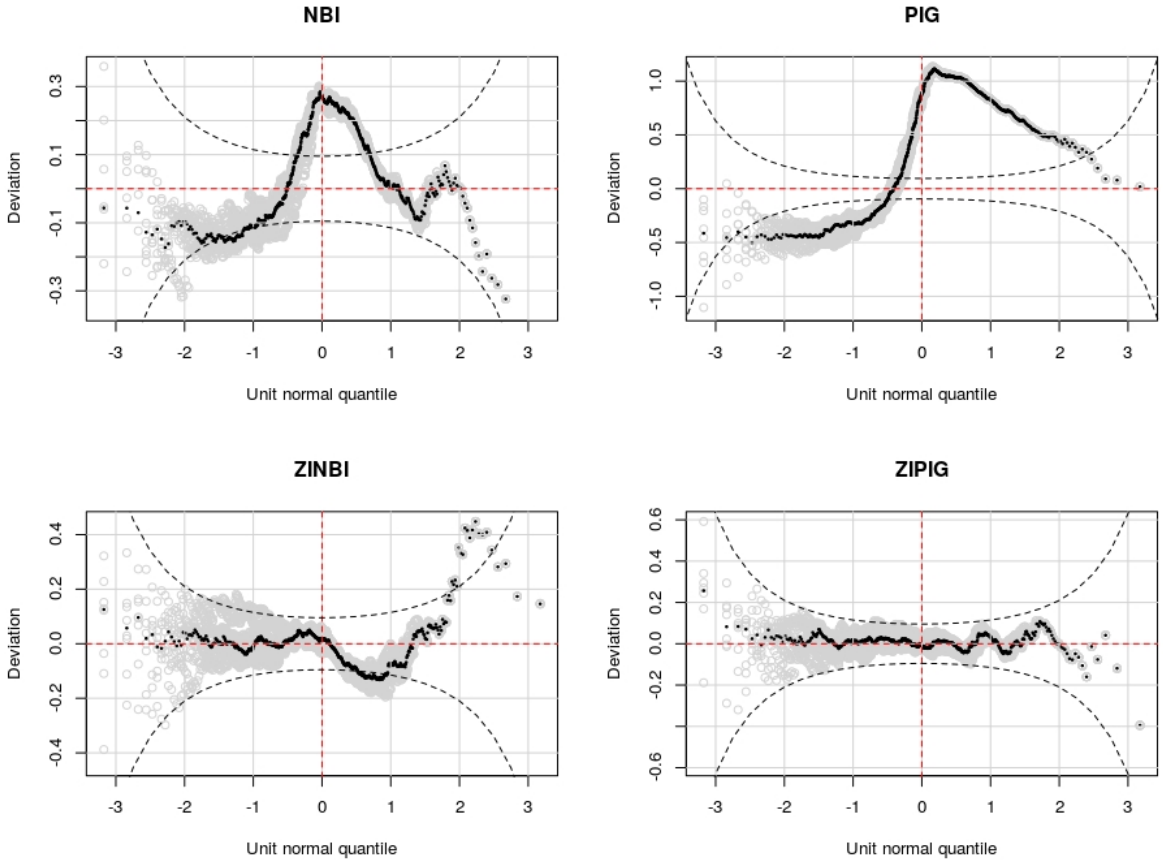
Distribuições	Preditores	Tratamentos	Faixa etária	Estação	Peso	Temperatura
NBI	μ	×	×	×	×	×
	σ	×	×	×	×	×
PIG	μ	×	×	×		
	σ				×	×
ZINBI	μ		×		×	×
	σ		×		×	×
	ν		×	×		
ZIPIG	μ		×		×	×
	σ					
	ν	×	×	×		

Na Figura 4.5, podemos observar os gráficos de minhoca dos resíduos dos modelos obtidos por meio da função `rqres.plot()`. Nos argumentos da função, definimos que seriam realizadas 8 aleatorizações[1], que são mostradas nos pontos cinzas, e os ponto pretos são a média deles. Podemos observar que o modelo ZIPIG obteve o melhor ajuste, pois é a única distribuição em que todos os pontos estão dentro das bandas de confiança, além de não apresentarem quaisquer padrões (linear, quadrático e/ou cúbico, como na

[1] Processo de aleatorização dos resíduos quantílicos normalizados para distribuições discretas apresentado na Seção 2.5.

Tabela 2.1). Consideramos então que tal modelo será o modelo final, e será analisado na Seção 4.3. O modelo binomial negativo (NBI) mostrou um formato acentuado em U invertido indicando que a assimetria foi superestimada. Já o modelo Poisson-normal inversa (PIG) mostrou um formato em S com curva esquerda para cima, indicando que a cauda da distribuição ajustada foi muito leve.

Figura 4.5 – Gráficos de minhoca para os modelos ajustados. Os pontos cinzas são as aleatorizações, e os pretos, a média delas.



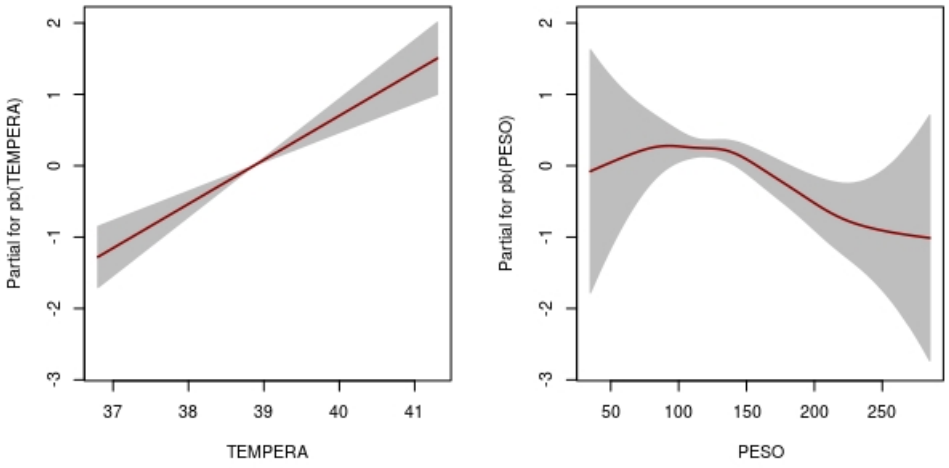
4.3 Modelo Poisson-normal inversa inflacionado de zeros (ZIPIG)

Utilizamos a função `term.plot()` para observar o relacionamento entre os parâmetros e as variáveis explicativas contínuas, que inicialmente ajustamos com a suavização *P-Spline* (`pb()`). Esta função nos retorna um gráfico que mostra os valores das variáveis explicativas *versus* os valores do parâmetro que à contém em seu preditor. O resultado pode ser visualizado na Figura 4.6.

A variável temperatura mostra uma relação linear com o parâmetro μ, portanto é desnecessário utilizar uma função de suavização. Assim retiramos a função *P-spline* de

temperatura em μ. Já com relação ao peso, podemos observar que o parâmetro μ cresce lentamente entre os pesos 50 e 100kg, se mantém entre 100 e 150kg, e começa a decrescer a partir de 150kg. Como o peso mostra uma relação não-linear, mantivemos a função de suavização.

Figura 4.6 – Saída da função `term.plot()` para o modelo ZIPIG.



O modelo final é mostrado na Equação 4.1 e as estimativas, erros padrões, estatística do teste de Wald e valor-p encontram-se na Tabela 4.5, sendo que estes podem ser obtidos por meio da função `summary()`. Consideramos um nível de 5% de significância. Cabe aqui relembrar o que os parâmetros do modelo ZIPIG representam. Os parâmetros μ e σ são a média e a dispersão do componente PIG, respectivamente. Já o parâmetro ν, refere-se à probabilidade extra da contagem ser nula.

$$\begin{aligned} \boldsymbol{OPG} &\overset{ind}{\sim} ZIPIG(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}) \\ \boldsymbol{\eta}_1 = ln(\boldsymbol{\mu}) &= \beta_{01} + pb(PESO) + TEMPERA + FAIXAETARIA \\ \boldsymbol{\eta}_2 = ln(\boldsymbol{\sigma}) &= \beta_{02} \\ \boldsymbol{\eta}_3 = logit(\boldsymbol{\nu}) &= \beta_{03} + FAIXAETARIA + TRAT + ESTACAO. \end{aligned} \tag{4.1}$$

Na Tabela 4.5, podemos observar que um nível do fator faixa etária está oculto, isto ocorre porque só se é possível obter estimativas para fatores, considerando alguma restrição (RENCHER; SCHAALJE, 2008). A restrição utilizada por padrão no *software* `R` é considerar que o primeiro nível é nulo, de forma que apresentamos os efeitos dos fatores estimáveis.

Tabela 4.5 – Resumo do modelo ZIPIG[a].

Coeficientes	Estimativa	Erro padrão	Estatística t	Valor-p
Preditor de μ com ligação log				
(Intercepto)	-17,050070	3,827240	-4,455	$9{,}88\times10^{-6}$
FAIXAETARIA181-365	1,556296	0,379538	4,100	$4{,}65\times10^{-5}$
FAIXAETARIA91-180	1,667091	0,335205	4,973	$8{,}43\times10^{-7}$
TEMPERA	0,618219	0,097681	6,329	$4{,}60\times10^{-10}$
pb(PESO)	–	–	–	–
Preditor de σ com ligação log				
(Intercepto)	0,8539	0,1163	7,342	6.31×10^{-13}
Preditor de ν com ligação logit				
(Intercepto)	2,7692	0,3420	8,096	$2{,}8\times10^{-15}$
FAIXAETARIA181-365	-3,1702	0,3098	-10,232	$< 2\times10^{-16}$
FAIXAETARIA91-180	-3,0722	0,3261	-9,421	$< 2\times10^{-16}$
TRATESTRAT	0,6000	0,1844	3,255	0,001195
ESTACAOOUTONO	-0,6361	0,2493	-2,552	0,010944
ESTACAOPRIMAVERA	-0,5809	0,2494	-2,329	0,020171
ESTACAOVERAO	-0,9795	0,2679	-3,656	0,000277

[a] Os interceptos são os parâmetros β_0 do modelo 4.1, FAIXAETARIA181-365 é o nível designado para a bezerra que se encontra na idade de 181 à 365 dias e FAIXAETARIA91-180 para aquelas entre 91 e 180 dias. TEMPERA contém os valores de temperatura e PESO, peso. TRATESTRAT designa o efeito do tratamento estratégico e ESTACAOOUTONO, ESTACAOPRIMAVERA e ESTACAOVERAO designam os fatores das estações do ano em que foram realizados as coletas.

Isolando os parâmetros μ e σ em 4.1, que resulta na inversa da função de ligação *log*, temos que

$$\begin{aligned} \mu = e^{\eta_1} = e^{\beta_{01}} \times e^{pb(PESO)} \times e^{TEMPERA} \times e^{FAIXAETARIA} \\ \sigma = e^{\eta_2} = e^{\beta_{02}}. \end{aligned} \tag{4.2}$$

Substituindo os coeficientes estimados (Tabela 4.5) na Equação 4.2, podemos quantificar os efeitos das covariáveis nos parâmetros. O intercepto de μ é equivalente à $3,94 \times 10^{-8}$, sendo que, se a bezerra está contida na faixa etária de 0-90 dias, não esperamos nenhuma alteração no valor deste parâmetro, porque, pela restrição utilizada, este nível é nulo em relação ao intercepto. Mas se contida na faixa etária de 91-180 dias, esperamos que μ aumente $5,31$ vezes em relação ao nível 0-90 dias e se contida em 181-365 dias, espera-se que μ aumente $4,81$ vezes em relação ao nível 0-90 dias. Agora, para cada aumento em uma unidade do valor da temperatura, espera-se que o parâmetro μ aumente $1,86$ vezes. A relação da covariável peso com o parâmetro μ foi discutida no início desta seção. Como ela contém a função de suavização *P-spline*, não possui estimativas pontuais.

Com relação ao parâmetro σ, equivalente à dispersão do componente PIG, foi ajustado uma constante, não sendo influenciado pelas covariáveis, ou seja, para quaisquer que sejam às características da bezerra, seu valor será sempre equivalente à $2,35$.

Já o parâmetro ν, que representa a probabilidade extra da contagem ser nula, temos a função inversa da ligação *logit*, isto é (CRAMER, 2003)

$$\boldsymbol{\eta}_3 = ln\left(\frac{\boldsymbol{\nu}}{1-\boldsymbol{\nu}}\right) \implies \boldsymbol{\nu} = \frac{e^{\boldsymbol{\eta}_3}}{1+e^{\boldsymbol{\eta}_3}},$$

em que $\boldsymbol{\eta}_3$ é dado em 4.1.

Considerando que uma bezerra recebeu o tratamento convencional e o dado foi coletado no inverno, então o valor de ν será $0,9404$ para a faixa etária 0-90 dias, $0,4231$ se 91-180 dias e $0,3989$ para 181-365 dias.

Agora, se a bezerra encontra-se na faixa etária 0-90 dias, e o dado foi coletado no inverno, então o valor de ν é $0,9404$, se recebeu o tratamento convencional e, $0,9664$ se recebeu o tratamento estratégico. Por fim, analisando o efeito de estação, se a bezerra está na faixa etária 0-90 dias e recebeu o tratamento convencional, com o dado coletado no inverno, o valor de ν é $0,9404$, no outono igual à $0,8932$, na primavera $0,8983$ e por fim, no verão, o valor de ν é $0,8557$.

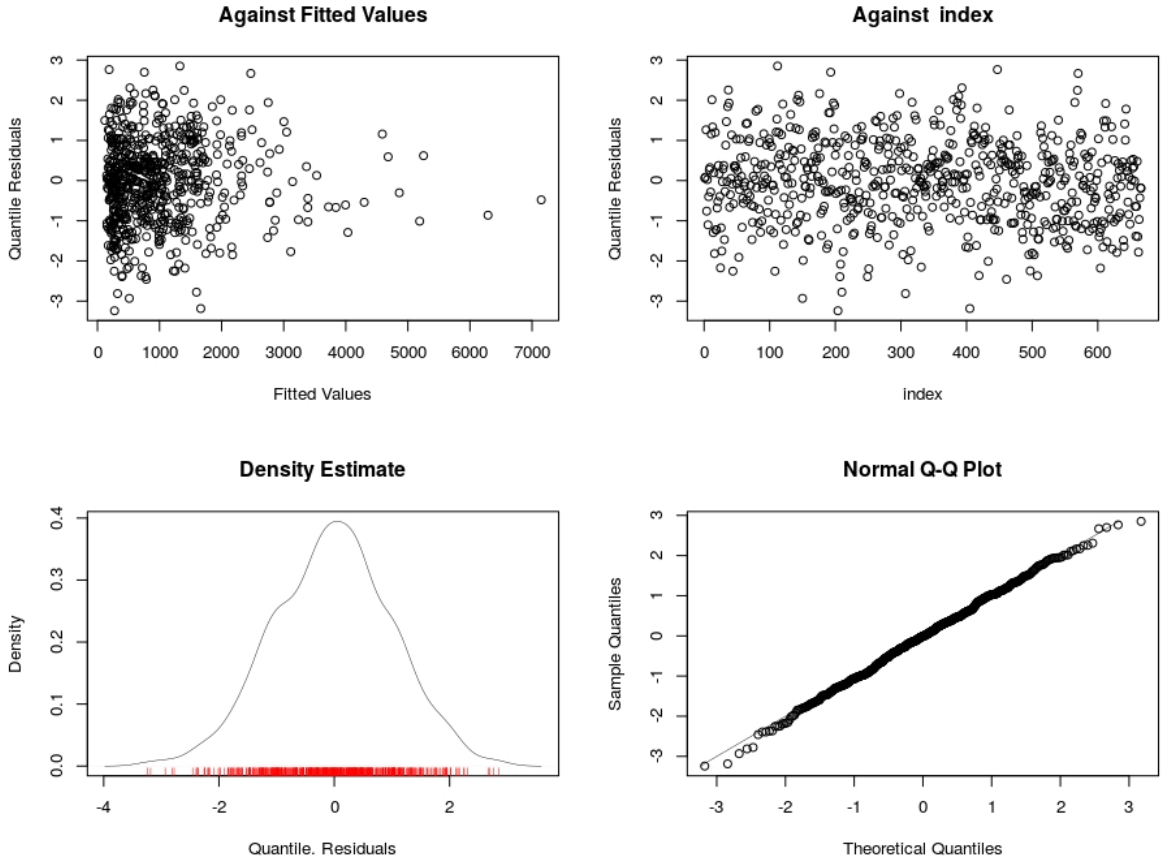
Muitas combinações podem ser feitas para obter a probabilidade extra da contagem nula. Em suma, pelos sinais dos coeficientes, podemos concluir que a probabilidade será maior para a faixa etária 0-90 dias, e menores nas demais faixas. Também o tratamento estratégico aumenta a probabilidade da contagem ser nula. Com relação às estações, verão apresenta menor probabilidade de contagem nula em relação às demais estações. A estação inverno é a que apresenta maior probabilidade de contagem nula.

A Figura 4.7 mostra a saída da função `plot()`. No gráfico de resíduos *versus* valores ajustados do parâmetro μ (*Against Fitted Values*), podemos observar que os pontos se distribuem aleatoriamente em torno dos resíduos nulos, o que mostra uma boa adequação. Também podemos observar um maior número de pontos em contagens baixas pela própria natureza da distribuição, com inflação em zeros.

Stasinopoulos e Rigby (2007), ao adequarem a função `plot()` para GAMLSS, substituem os conhecidos valores ajustados pelos valores ajustados dos parâmetros de locação da distribuição de cada observação, porque em GAMLSS não existe valores ajustados pontuais, como em ajustes de modelos lineares, agora temos em mãos toda a distribuição

ajustada para aquela observação. No gráfico de resíduos *versus* indexação da ordem das observações (*Against index*), podemos perceber uma nuvem de pontos completamente aleatória, sendo um resultado desejável que confirma a homocedasticidade dos resíduos. No gráfico da estimação da densidade kernel dos resíduos (*Density Estimate*), podemos observar um formato semelhante à um sino que mostra que os resíduos são normalmente distribuídos. Este diagnóstico é reafirmado a partir do gráfico normal Q-Q dos resíduos (*Normal Q-Q Plot*), em que os pontos estão distribuídos como uma diagonal, com poucos pontos fora dela na parte inferior e superior da distribuição.

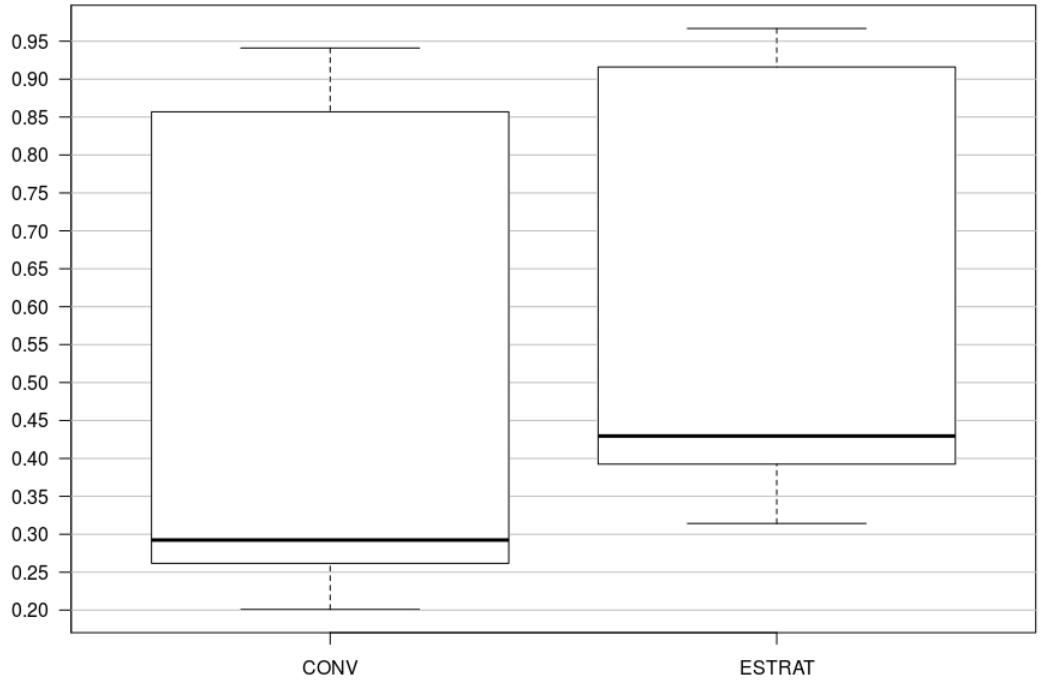
Figura 4.7 – Saída da função `plot()`.



A partir do modelo final, dado pela Equação 4.1, podemos ter mais alguns resultados. O experimento de Blanco (2015) foi realizado com intuito de verificar o efeito de um novo tratamento contra doenças parasitárias em bezerras, de forma que focaremos as análises no estudo de tal efeito. Em GAMLSS, obtemos uma função (densidade) de probabilidade para cada observação, com diferentes valores para os parâmetros à medida que os valores das covariáveis se alteram. Na Figura 4.8, temos um gráfico de caixas estratificado que sintetiza o comportamento de ν para os tratamentos convencional e estratégico obtido

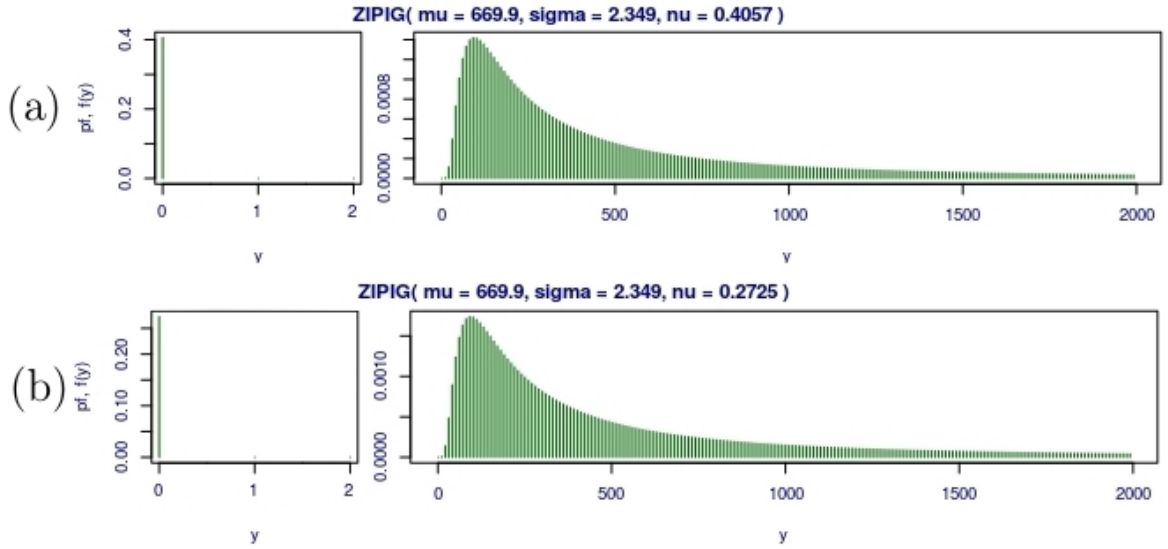
por cada observação da amostra. Podemos observar que, para o tratamento estratégico, os valores de ν são consideravelmente superiores em relação ao convencional, mostrando a efetividade do tratamento estratégico em detrimento do convencional, porque maiores valores de ν significam maiores probabilidades de contagens OPG nulas, que ocorrem quando a bezerra está saudável.

Figura 4.8 – Valores preditos de ν para cada tratamento.



Além disso, podemos realizar predições e verificar o efeito de tratamentos pontualmente por meio da função `predictAll()`. Suponha uma bezerra, chamada Baroninha, classificada na faixa etária de 181 dias a 1 ano, com temperatura corporal de $38,7^oC$ e peso de 190kg e que a estação do ano seja primavera. Escolhemos estes valores por serem os valores médios observados na amostra. É importante ressaltar que os valores a serem preditos precisam estar dentro do intervalo observado da amostra, para obtermos estimativas confiáveis. Neste caso, se Baroninha estiver recebendo o tratamento convencional, então $\nu = 0,2725$. Porém, se o tratamento for o estratégico, então $\nu = 0,4057$. Assim, existe um aumento aproximado de 48,9% da probabilidade da contagem OPG desta bezerra ser nula se utilizado o tratamento estratégico. Na Figura 4.9, podemos observar a função de probabilidade obtida pela função `pdf.plot()` para a Baroninha, caso tenha recebido o tratamento estratégico (a) ou convencional (b).

Figura 4.9 – Função de probabilidade para a contagem de OPG esperada para a Baroninha.



Como já mencionado, o efeito de tratamentos se alterará à medida que as características das bezerras forem diferentes. Na Figura 4.9, as probabilidades de contagens nulas foram divididas nos gráficos para melhor visualização, uma vez que a probabilidade da contagem nula é bem discrepante em relação as demais, bastando observar os valores do eixo das ordenadas `pf,f(y)`. Também limitamos o gráfico para contagens até 2000, porque valores mais altos possuem probabilidade ínfima. Podemos notar também que o parâmetro μ se alterou em função das características das covariáveis, uma vez que, no modelo ajustado, este parâmetro depende da faixa etária, temperatura e peso. Como já apresentado na Seção 3.2, para uma distribuição ZIPIG, a esperança e a variância são dadas por

$$E(Y) = (1 - \nu)\mu \tag{4.3}$$

$$Var(Y) = \mu(1 - \nu) + \mu^2(1 - \nu)(\sigma + \nu). \tag{4.4}$$

Assim, podemos calcular estes momentos para Baroninha, bastando substituir os parâmetros estimados nas Equações 4.3 e 4.4. Os valores obtidos são exibidos na Tabela 4.6.

Podemos observar que tanto a média quanto o desvio padrão foram menores para o tratamento estratégico. Isto sempre ocorrerá porque o parâmetro ν influencia de forma inversamente proporcional no cálculo da média e da variância de uma distribuição ZIPIG.

Tabela 4.6 – Estimativas da média, variância e desvio padrão da contagem de OPG para Baroninha.

	Convencional	Estratégico
Média	487,53	398,12
Variância	856347,50	735081,10
Desvio Padrão	925,39	857,36

Em outra palavras, o tratamento estratégico aumenta o valor de ν em relação ao convencional, ou seja, o tratamento estratégico aumenta a probabilidade da contagem ser nula, o que nos levará a menores valores médios da contagem OPG.

5 CONSIDERAÇÕES FINAIS

- **Flexibilidade e boa adequação dos GAMLSS:** O modelo GAMLSS aparentou ser mais apropriado para descrever a contagem OPG, uma vez que esta é uma variável que apresenta excesso de zeros, permitindo distribuições próprias para descrever tal característica, como as distribuições de inflação de zeros. O modelo também foi vantajoso, pois, não só o parâmetro de locação da distribuição ZIPIG dependia de covariáveis, mas também o parâmetro de forma. Assim, descrevemos e interpretamos a natureza desta variável de uma maneira bem mais realística.

- **Riqueza de informações:** Os GAMLSS permitem a obtenção de uma distribuição de probabilidade para cada observação. No caso, temos uma distribuição de probabilidade da contagem OPG para cada bezerra e, à medida que mudam suas características, a distribuição também mudará, o que certamente é útil na prática, trazendo informações mais completas para análise.

- **Predição:** Além de obter funções de probabilidade para cada situação amostrada, podemos obter uma distribuição referente a qualquer situação que englobe os valores e situações definidas nas variáveis explicativas, mesmo que não amostrada, contanto que não haja extrapolações.

- **Vantagem:** Os modelos GAMLSS se mostram promissores para análise de dados univariados, não só em experimentos agropecuários, como em diversas áreas, porque conseguem unificar, em uma só teoria, vários procedimentos estatísticos.

5 REFERÊNCIAS

ABBAS, M. et al. Biodiversity effects on plant stoichiometry. **PLoS One**, Public Library of Science, v. 8, n. 3, p. e58179, 2013.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: **Selected Papers of Hirotugu Akaike**. [S.l.]: Springer, 1998. p. 199–213.

BARAJAS, F. H. et al. Gamlss models applied in the treatment of agro-industrial waste. **Comunicaciones en Estadística**, v. 8, n. 2, p. 245–254, 2015.

BLANCO, Y. A. C. **Efeito e custos do tratamento estratégico seletivo no controle de parasitoses gastrointestinais em bezerras leiteiras**. Dissertação (Mestrado) — Departamento de Zootecnia da Universidade Federal de Lavras, 2015.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in medicine**, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.

CHARNET, R. et al. **Análise de Regressão Linear: com aplicações**. [S.l.]: Editora da UNICAMP, 2008. v. 2ª edição.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. **Statistics in medicine**, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992.

CRAMER, J. S. The origins and development of the logit model. **Logit models from economics and other fields**, Citeseer, v. 2003, p. 1–19, 2003.

DEAN, C.; LAWLESS, J.; WILLMOT, G. A mixed poisson–inverse-gaussian regression model. **Canadian Journal of Statistics**, Wiley Online Library, v. 17, n. 2, p. 171–181, 1989.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. [S.l.]: USP/ESALQ, 2001.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.

EILERS, P. H.; MARX, B. D.; DURBÁN, M. Twenty years of p-splines. **SORT: statistics and operations research transactions**, v. 39, n. 2, p. 0149–186, 2015.

ELOY, R. A. O. **Distribuição assimétrica t-student tipo 3: uma aplicação a delineamentos inteiramente casualizados**. 2016. Monografia (Especialização em Estatística Aplicada) - Universidade Estadual da Paraíba, Campina Grande, Brazil.

HASTIE, T.; TIBSHIRANI, R. **Generalized additive models**. [S.l.]: Wiley Online Library, 1990.

HASTIE, T. J. Generalized additive models. In: **Statistical models in S**. [S.l.]: Routledge, 2017. p. 249–307.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. **Univariate discrete distributions**. [S.l.]: John Wiley & Sons, 2005. v. 444.

KOMSTA, L.; NOVOMESTKY, F. **moments: Moments, cumulants, skewness, kurtosis and related tests**. [S.l.], 2015. R package version 0.14. Disponível em: <https://CRAN.R-project.org/package=moments>.

LAMBERT, D. Zero-inflated poisson regression with an application to defects in manufacturing. **Technometrics**, Taylor & Francis, v. 34, n. 1, p. 1–14, 1992.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. [S.l.]: CRC press, 1989. v. 37.

NAKAMURA, L. R. et al. Modelling location, scale and shape parameters of the birnbaum-saunders generalized t distribution. **Journal of Data Science**, v. 15, n. 2, p. 221–237, 2017.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, v. 135, n. 3, p. 370–384, 1972.

PEIXOTO, M. d. M.; SAGE, R. F. Improved experimental protocols to evaluate cold tolerance thresholds in miscanthus and switchgrass rhizomes. **Gcb Bioenergy**, Wiley Online Library, v. 8, n. 2, p. 257–268, 2016.

PIEKARSKA-BONIECKA, H. et al. Model for bionomy of privet sawfly (macrophya punctumalbum l.)(hymenoptera, tenthredinidae). **Acta Scientiarum Polonorum. Hortorum Cultus**, -, v. 9, n. 3, 2010.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

RENCHER, A. C.; SCHAALJE, G. B. **Linear models in statistics**. [S.l.]: John Wiley & Sons, 2008.

RIGBY, R. et al. **Distributions for Modelling Location, Scale, and Shape: Using GAMLSS in R**. [S.l.: s.n.], 2017.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005.

RIGBY, R. A.; STASINOPOULOS, D. M. Automatic smoothing parameter selection in gamlss with an application to centile estimation. **Statistical methods in medical research**, Sage Publications Sage UK: London, England, v. 23, n. 4, p. 318–332, 2014.

RIGHETTO, A. J. et al. Predicting weed invasion in a sugarcane cultivar using multispectral image. **Journal of Applied Statistics**, Taylor & Francis, p. 1–12, 2018.

ROCHA, L. O. et al. Mycoflora and co-occurrence of fumonisins and aflatoxins in freshly harvested corn in different regions of brazil. **International journal of molecular sciences**, Molecular Diversity Preservation International, v. 10, n. 11, p. 5090–5103, 2009.

RODGERS, J. L.; NICEWANDER, W. A. Thirteen ways to look at the correlation coefficient. **The American Statistician**, Taylor & Francis, v. 42, n. 1, p. 59–66, 1988.

SCHWARZ, G. et al. Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance components**. [S.l.]: John Wiley & Sons, 2009. v. 391.

STASINOPOULOS, D. M.; RIGBY, R. A. Generalized additive models for location scale and shape (gamlss) in r. **Journal of Statistical Software**, v. 23, n. 7, p. 1–46, 2007.

STASINOPOULOS, M. D. et al. **Flexible Regression and Smoothing: Using GAMLSS in R**. [S.l.]: CRC Press, 2017.

VOUDOURIS, V. et al. Modelling skewness and kurtosis with the bcpe density in gamlss. **Journal of Applied Statistics**, Taylor & Francis, v. 39, n. 6, p. 1279–1293, 2012.

ZEVIANI, W. M.; JR, E. E. R.; TACONELI, C. A. Modelos de regressão para dados de contagem com r. **Anais da 61ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria.**, 2016.

A APÊNDICES

Código em R utilizado na aplicação

```
#-------------------------------------------------------------------------
#Script da dissertacao: GAMLSS na experimentacao agropecuaria
#Autora: Fernanda Venturato Roquim
#Data: 18 de maio de 2018

library(gamlss)
library(moments)
dat <-read.csv('~/OPG/bezerro.csv')
attach(dat)
#-------------------------------------------------------------------------
#Analise exploratoria da contagem OPG.

plot(OPG, xlab ="Observao")
mean(OPG)
sd(OPG)
boxplot(OPG)
summary(OPG)
skewness(OPG)
kurtosis(OPG)

#Graficos de dispersao.
par(mfrow=c(2,3))
plot(PESO, OPG, main ="Peso")
plot(ALTURA,OPG, main ="Altura")
plot(TEMPERA, OPG, main ="Temperatura")
plot(TRAT, OPG, main= "Tratamentos")
plot(FAIXAETARIA, OPG, main="Faixa etria")
plot(ESTACAO, OPG, main= "Estao do ano")
#-------------------------------------------------------------------------
#Analise exploratoria das variveis continuas.
```

```
31
32 summary(PESO)
33 summary(ALTURA)
34 summary(TEMPERA)
35      var(PESO); var(ALTURA); var(TEMPERA)
36       sd(PESO); sd(ALTURA); sd(TEMPERA)
37 skewness(PESO); skewness(ALTURA); skewness(TEMPERA)
38 kurtosis(PESO); kurtosis(ALTURA); kurtosis(TEMPERA)
39
40 par(mfrow=c(1,3))
41 boxplot(PESO, xlab ="Peso")
42 boxplot(ALTURA, xlab ="Altura")
43 boxplot(TEMPERA, xlab ="Temperatura")
44
45 plot(PESO, ALTURA)
46 cor(PESO, ALTURA)
47 #-----------------------------------------------------------------------------
48 #Analise exploratoria dos fatores.
49
50 summary(TRAT)
51 summary(FAIXAETARIA)
52 summary(ESTACAO)
53
54 par(mfrow=c(1,1))
55
56 trat <-table(TRAT)
57 pie(trat, main ="Tratamentos")
58 barplot(trat, main ="Tratamentos", ylab= "Nmero de observaes" )
59 prop.table(trat)
60
61 faixa <-table(FAIXAETARIA)
62 pie(faixa, main ="Faixa etria em dias")
63 barplot(faixa, main ="Faixa etria em dias")
64 prop.table(faixa)
```

```
65
66 esta <-table(ESTACAO)
67 pie(esta, main ="Estao")
68 barplot(esta, main ="Estao")
69 prop.table(esta)
70 #-------------------------------------------------------------------------
71 #Associacao entre variaveis resposta e explicativas.
72
73 plot(PESO, OPG)
74 plot(TEMPERA, OPG)
75 cor(OPG, PESO)
76 cor(OPG, TEMPERA)
77
78 boxplot(OPG, TRAT)
79 boxplot(OPG, FAIXAETARIA)
80 #-------------------------------------------------------------------------
81 #Distribuicao marginal para a resposta
82
83 a <-fitDist(dat$OPG, type='counts')
84 a$fits
85 #-------------------------------------------------------------------------
86 #Modelo Poisson
87
88 m1 <-gamlss(OPG~1, data=dat, family=PO)
89 mod1 <-stepGAICAll.A(m1, scope=list(lower=~1,
90                                     upper=~TRAT+FAIXAETARIA+ESTACAO+pb(PESO)+pb(
                                        TEMPERA)))
91 summary(mod1)
92 #-------------------------------------------------------------------------
93 #Modelo Binomial Negativo
94
95 m2 <-gamlss(OPG~1, data=dat, family=NBI)
96 mod2 <-stepGAICAll.A(m2, scope=list(lower=~1,
```

```
97                                  upper=~TRAT+FAIXAETARIA+ESTACAO+pb(PESO)+pb(
                                        TEMPERA)))
98  summary(mod2)
99  #-------------------------------------------------------------------------
100 #Modelo Poisson Normal Inversa
101
102 m3 <-gamlss(OPG~1, data=dat, family=PIG, method=mixed(30,500))
103 mod3 <-stepGAICAll.A(m3, scope=list(lower=~1,
104                                  upper=~TRAT+FAIXAETARIA+ESTACAO+pb(PESO)+pb(
                                        TEMPERA)))
105 summary(mod3)
106 #-------------------------------------------------------------------------
107 #Modelo Poisson Zero Inflado
108
109 m4 <-gamlss(OPG~1, data=dat, family=ZIP, method=mixed(10,100),
110             gd.tol=Inf)
111 mod4 <-stepGAICAll.A(m4, scope=list(lower=~1,
112                                  upper=~TRAT+FAIXAETARIA+ESTACAO+pb(PESO)+pb(
                                        TEMPERA)))
113 summary(mod4)
114 #-------------------------------------------------------------------------
115 #Modelo Binomial Nagativo Zero Inflado
116
117 m5 <-gamlss(OPG~1, data=dat, family=ZINBI , method=mixed(10,100),
118             gd.tol=Inf)
119 mod5 <-stepGAICAll.A(m5, scope=list(lower=~1,
120                                  upper=~TRAT+FAIXAETARIA+ESTACAO+pb(PESO)+pb(
                                        TEMPERA)))
121 summary(mod5)
122 #-------------------------------------------------------------------------
123 #Modelo Poisson Normal Inversa Zero Inflado
124
125 m6 <-gamlss(OPG~1, data=dat, family=ZIPIG, method=mixed(15,100))
126 mod6 <-stepGAICAll.A(m6, scope=list(lower=~1,
```

```
127                                    upper=~TRAT+FAIXAETARIA+ESTACAO+pb(PESO)+pb(
                                           TEMPERA)))
128 summary(mod6)
129
130 #retirando TRAT de mu porque no foi significativo
131 modelo6 <-gamlss(formula =OPG ~FAIXAETARIA +pb(TEMPERA) +
132                    pb(PESO), sigma.formula =~1, nu.formula =~FAIXAETARIA +
133                    TRAT +ESTACAO, family =ZIPIG, data =dat, method =RS(100),
134                  trace =TRUE)
135 summary(modelo6)
136 #-------------------------------------------------------------------------
137 #Relacionamento entre os parametros e as variaveis explicativas continuas.
138
139 par(mfrow =c(1,1))
140 term.plot(modelo6, what='mu') #relacionamento entre as variaveis explicativas
        continuas e \mu
141 #-------------------------------------------------------------------------
142 #Definindo o modelo final.
143
144 modelofinal <-gamlss(formula =OPG ~FAIXAETARIA +TEMPERA +
145                       pb(PESO), sigma.formula =~1, nu.formula =~FAIXAETARIA +
146                       TRAT +ESTACAO, family =ZIPIG, data =dat, method =RS(100),
147                       trace =TRUE)
148 summary(modelofinal)
149 par(mfrow=c(2,2))
150 plot(modelofinal, parameters =mu)
151 plot(modelofinal, ts=T)
152
153 #-------------------------------------------------------------------------
154 #Obtendo os graficos de minhoca com rqres.plot().
155
156 par(mfrow=c(2,2))
157 rqres.plot(mod2, howmany =8, cex=.5, pch=20, col='black', ylim.all =2, plot.type
        ="all"); title("NBI")
```

```
158 rqres.plot(mod3, howmany =8, cex=.5, pch=20, col='black', ylim.all =2, plot.type
        ="all"); title("PIG")
159 rqres.plot(mod5, howmany =8, cex=.5, pch=20, col='black', ylim.all =2, plot.type
        ="all"); title("ZINBI")
160 rqres.plot(modelofinal, howmany =8, cex=.5, pch=20, col='black', ylim.all =2,
       plot.type ="all"); title("ZIPIG")
161 #------------------------------------------------------------------------
162 #Calculando os valores dos parametros (inverso da funcao de ligacao)
163 #Substitua os valores pelos coeficientes estimados.
164
165 #mu
166 exp(-17.05) #intercepto
167 exp(1.67) #faixa etaria 91-180
168 exp(1.57) #faixa etria 181-365
169 exp(0.62) #temperatura
170
171 #sigma
172 exp(0.8539) #intercepto
173
174 #nu
175 #efeito de faixa etria fixando trat convencional estao inverno
176 eta =2.76 #se 0-90
177 eta =2.76 -3.07 #se 91-180
178 eta =2.76 -3.17 #se 181-365
179
180 #efeito de trat fixando 0-90 e inverno
181 eta =2.76 #se trat convencional
182 eta =2.76 +0.6 #se trat estratgico
183
184 #efeito de estao fixando 0-90 e convencional
185 eta =2.76 #se inverno
186 eta =2.76 -0.6361 #se outono
187 eta =2.76 -0.5809 #se primavera
188 eta =2.76 -0.9795 #se vero
```

```
189
190 nu =exp(eta)/(1+exp(eta)) ;nu
191 #-------------------------------------------------------------------------
192 #Obtendo predicoes
193
194 predict <-predictAll(modelofinal)
195 predict(modelofinal)
196 par(mfrow=c(1,1))
197 plot(dat$TRAT, predict$nu, yaxt ="n" )
198 axis(side=2, at=seq(0, 1, by=0.05), las =1)
199 for (a in seq(0, 1, by=0.05))
200 {abline(a, b=0, col ="lightgray")}
201
202 preditosnu <-data.frame(dat$TRAT, predict$nu)
203 summary(preditosnu)
204 tapply(predict$nu, dat$TRAT, mean)
205 #-------------------------------------------------------------------------
206 #Estimativas para a Baroninha
207
208 novosdados1 <-data.frame(FAIXAETARIA= c("181-365","181-365"),
209                          TEMPERA= c(38.7, 38.7),
210                          PESO= c(190, 190),
211                          TRAT= c("ESTRAT", "CONV"),
212                          ESTACAO= c("PRIMAVERA", "PRIMAVERA"))
213
214 predict1 <-predictAll(modelofinal, newdata =novosdados1)
215
216 pdf.plot(mu=predict1$mu, sigma=predict1$sigma, nu=predict1$nu, family=ZIPIG,
217          min=0, max= 2, step=1)
218
219 pdf.plot(mu=predict1$mu, sigma=predict1$sigma, nu=predict1$nu, family=ZIPIG,
220          min=1, max= 2000, step=10)
221 #-------------------------------------------------------------------------
222 #Obtendo media, variancia e desvio padrao (substitua os valores).
```

```
mu =382.8
si =2.349
nu =0.916

var =mu*(1-nu) +mu^2* (1-nu)*(si +nu) ;var
sqrt(var)
#-------------------------------------------------------------------------
```

ScriptFinal.R

B DISTRIBUIÇÕES

Distribuições contínuas em $(-\infty, \infty)$

- *Distribuições com dois parâmetros:*

  Gumbel: Esta distribuição é apropriada para dados moderadamente assimétricos à esquerda; Notação: `GU(`μ, σ`)`; Função de ligação padrão: identidade para μ e log para σ.

  Logística: Esta distribuição é apropriada para dados com problemas leves de curtose; Notação: `LO(`μ, σ`)`; Função de ligação padrão: identidade para μ e log para σ.

  Normal: Esta distribuição é padrão da função `gamlss()` se nenhuma família for definida; Notação: `NO(`μ, σ`)`; Função de ligação padrão: identidade para μ e log para σ.

  Gumbel reversa: Também conhecida como distribuição de valores extremos tipo I, apropriada para dados moderadamente assimétricos à direita; Notação: `RG(`μ, σ`)`; Função de ligação padrão: identidade para μ e log para σ.

- *Distribuições com três parâmetros:*

  Exponencial gaussiana: também conhecida como normal *lagged*. Se adéqua à dados assimétricos; Notação: `exGAUS(`μ, σ, ν`)`; Função de ligação padrão: identidade para μ, log para σ e log para ν.

  Exponencial potência: Tem como casos particulares as distribuições Normal e Laplace. Utilizada para dados platicúrticos ou leptocúrticos; Notação: `PE(`μ, σ, ν`)`; Função de ligação padrão: identidade para μ, log para σ e log para ν.

  Skew normal tipo I: Modela assimetria. As distribuições *skew* possuem uma cauda da distribuição mais longa que a outra; Notação: `SN1(`μ, σ, ν`)`; Função de ligação padrão: identidade para μ, log para σ e identidade para ν.

  Skew normal tipo II: Uma variação da tipo I, também modela assimetria; Notação: `SN2(`μ, σ, ν`)`; Função de ligação padrão: identidade para μ, log para σ e log para ν.

Família t: Apropriada para dados leptocúrticos, com curtose mais alta que a Normal; Notação: `TF`(μ, σ, ν); Função de ligação padrão: identidade para μ, log para σ e log para ν.

- *Distribuições com quatro parâmetros:*

  Exponencial beta generalizada tipo II: Também conhecida como logística generalizada tipo VI; Notação: `EGB2`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e log para τ.

  t generalizada: seus dois parâmetros de forma estão associados à curtose. É uma distribuição simétrica; Notação: `GT`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e log para τ.

  Johnson SU original: é leptocúrtica e modela assimetria e curtose; Notação: `JSUo`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

  t-normal-exponencial: apesar de conter quatro parâmetros, GAMLSS os parâmetros ν e τ são constantes. É simétrica; Notação: `NET`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ.

  Sinh-Arcsinh: modela caudas pesadas à direita e à esquerda; Notação: `SHASH`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e log para τ.

  Sinh-Arcsinh original: modela assimetria e curtose; Notação: `SHASHo`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

  Sinh-Arcsinh original tipo II: reparametrização de `SHASHo`; Notação: `SHASHo2`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

  Skew exponencial potência tipo I: modela assimetria e curtose. Tem como caso especial a *Skew* normal tipo I; Notação: `SEP1`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

Skew exponencial potência tipo II: também modela assimetria e curtose. Tem como casos especiais a *Skew* normal tipo I e a Normal; Notação: `SEP2`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

Skew exponencial potência tipo III: é uma distribuição *spliced-scale* modela a escala em duas partes. Tem como caso especial a *Skew* normal tipo II; Notação: `SEP3`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e log para τ.

Skew exponencial potência tipo IV: é uma distribuição *spliced-shape*, modela a forma em 2 partes. Notação: `SEP4`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e log para τ.

Skew t de Student: é uma reparametrização de `ST3`. Modela assimetria e curtose. Seu parâmetro σ é o próprio desvio padrão da distribuição; Notação: `SST`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e $\log(\tau - 2)$ para τ.

Skew t tipo I: a distribuição é diferente para y menor e maior que a média; Notação: `ST1`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

Skew t tipo II: modela assimetria e curtose. Esta distribuição é o caso univariado da *Skew t* multivariada; Notação: `ST2`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

Skew t tipo III: é uma distribuição *spliced-scale*, útil para quando existe interesse na moda, porque o parâmetro μ é a própria moda da distribuição; Notação: `ST3`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e log para τ.

Skew t tipo IV: é uma distribuição *spliced-shape*, seus parâmetros de forma estão associados às caudas esquerda e direita da distribuição; Notação: `ST4`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, log para ν e log para τ.

Skew t tipo V: modela assimetria e curtose; Notação: `ST5`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

Distribuições contínuas em $(0, \infty)$

- *Distribuição com um parâmetro:*

  Exponencial: única distribuição uniparamétrica contínua do pacote gamlss; Apropriada para dados levemente assimétricos à direita. Notação: `EXP(`μ`)`; Função de ligação padrão: log para μ.

- *Distribuições com dois parâmetros:*

  Gama: apropriada para dados com assimetria positiva; Notação: `GA(`μ, σ`)`; Função de ligação padrão: log para μ, log para σ.

  Gama Inversa: é um caso particular da gama generalizada. Um particularidade dela é que o parâmetro de escala é o μ, que também é a própria moda da distribuição; Notação: `IGAMMA(`μ, σ`)`; Função de ligação padrão: log para μ, log para σ.

  Normal Inversa: apropriada para dados fortemente assimétricos à direita. μ é a média das distribuição; Notação: `IG(`μ, σ`)`; Função de ligação padrão: log para μ, log para σ.

  Log-normal: apropriada para dados assimétricos à direita; Notação: `LOGNO(`μ, σ`)`; Função de ligação padrão: identidade para μ, log para σ.

  Pareto original tipo II: o parâmetro de escala é μ e não tem parâmetro de locação; Notação: `PARETO2o(`μ, σ`)`; Função de ligação padrão: log para μ, log para σ.

  Pareto tipo II: reparametrização de `PARETO2o`; Notação: `PARETO2(`μ, σ`)`; Função de ligação padrão: log para μ, log para σ.

  Weibull: possui três parametrizações em GAMLSS. É muito utilizadas em estudos de tempo de vida de produtos com taxas de falha; Notação: `WEI(`μ, σ`)`; Função de ligação padrão: log para μ, log para σ.

- *Distribuições com três parâmetros:*

  Box-Cox Cole e Green: apropriada para dados assimétricos à direita e à esquerda. Não possui média e variância; Notação: `BCCG(`μ, σ, ν`)`; Função de ligação padrão: identidade para μ, log para σ, e identidade para ν.

Gama generalizada: possui duas parametrizações em GAMLSS. É assimétrica e μ é parâmetro de escala; Notação: `GG`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ, e identidade para ν.

Normal inversa generalizada: tem como caso particular a normal inversa, mas modela a assimetria; Notação: `GIG`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ, e identidade para ν.

Família log normal: é resultado da aplicação da transformação Box-Cox potência em Y em uma normal. Modela assimetria. o parâmetro ν é constante; Notação: `LNO`(μ, σ, ν); Função de ligação padrão: identidade para μ, log para σ.

- *Distribuições com quatro parâmetros:*

Box-Cox t: modela assimetria e curtose. μ é o parâmetro de escala e σ é o coeficiente de variação aproximado; Notação: `BCT`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

Box-Cox exponencial potência: também modela assimetria e curtose. Também μ é o parâmetro de escala e σ é o coeficiente de variação aproximado; Notação: `BCPE`(μ, σ, ν, τ); Função de ligação padrão: identidade para μ, log para σ, identidade para ν e log para τ.

Beta generalizada tipo II: modela assimetria e curtose. Tem como casos particulares as distribuições Pearson tipo IV, Singh-Maddala e pareto tipo II; Notação: `GB2`(μ, σ, ν, τ); Função de ligação padrão: log para μ, log para σ, log para ν e log para τ.

Distribuições contínuas de mistura em $[0, \infty)$

Zero ajustada gama: adequada quando existe muitas observações nulas e as demais seguem uma gama. O parâmetro ν é a probabilidade exata de $Y = 0$; Notação: `ZAGA`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e logit para ν.

Zero ajustada normal inversa: adequada quando existe muitas observações nulas e as demais seguem uma normal inversa. O parâmetro ν é a probabilidade exata de $Y = 0$; Notação: `ZAIG`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e logit para ν.

Distribuições contínuas $(0, 1)$

- *Distribuições com dois parâmetros:*

  Beta: apropriada para dados com uma amplitude conhecida e restrita, excluindo os pontos finais do intervalo; Notação: `BE`(μ, σ); Função de ligação padrão: logit para μ, logit para σ.

- *Distribuições com quatro parâmetros:*

  Beta generalizada tipo I: modela assimetria e curtose, como a beta generalizada tipo II, mas para o intervalo 0 e 1. Notação: `GB1`(μ, σ, ν, τ); Função de ligação padrão: logit para μ, logit para σ, log para ν e log para τ.

Distribuições contínuas de mistura incluindo 0**,** 1 **ou ambos**

Beta Inflacionada (em 0 e 1): apropriada para dados com uma amplitude conhecida e restrita, incluindo os pontos finais do intervalo, com inflação de valores nos extremos da distribuição; Notação: `BEINF`(μ, σ, ν, τ); Função de ligação padrão: logit para μ, logit para σ, log para ν e log para τ.

Beta Inflacionada (em 0): apropriada para dados com uma amplitude conhecida e restrita, incluindo o ponto inicial do intervalo, com inflação de valores em 0; Notação: `BEINF0`(μ, σ, ν); Função de ligação padrão: logit para μ, logit para σ, log para ν.

Beta Inflacionada (em 1): apropriada para dados com uma amplitude conhecida e restrita, incluindo o ponto final do intervalo, com inflação em valores em 1; Notação: `BEINF1`(μ, σ, ν); Função de ligação padrão: logit para μ, logit para σ, log para ν.

Distribuições discretas de contagem

- *Distribuições com um parâmetro:*

  Geométrica: o parâmetro é a média da distribuição. Utilizada no estudo do número de falhas até o sucesso; Notação: `GEOM`(μ); Função de ligação padrão: log para μ.

  Logarítmica: o parâmetro μ recebe valores entre 0 e 1; Notação: `LG`(μ); Função de ligação padrão: log para μ.

Poisson: a principal característica é que μ é a média e a variância da distribuição; Notação: `PO(`μ`)`; Função de ligação padrão: log para μ.

Yule: é um caso particular da Waring. Esta distribuição tem caudas pesadas, principalmente para valores de μ altos. Notação: `YULE(`μ`)`; Função de ligação padrão: log para μ.

Zipf: também conhecida com pareto discreta. Se adéqua para dados com contagem com caudas muito pesadas; Notação: `ZIPF(`μ`)`; Função de ligação padrão: log para μ.

- *Distribuições com dois parâmetros:*

Poisson dupla: é um caso especial da exponencial dupla e tem como caso particular a Poisson quando $\sigma = 1$. É uma distribuição Poisson com superdispersão; Notação: `DPO(`μ, σ`)`; Função de ligação padrão: log para μ e log para σ.

Poisson generalizada: aproxima-se da binomial negativa; Notação: `GPO(`μ, σ`)`; Função de ligação padrão: log para μ e log para σ.

Binomial negativa: indicada para quando a média é diferente da dispersão; Notação: `NBI(`μ, σ`)`; Função de ligação padrão: log para μ e log para σ.

Poisson normal inversa: tem as mesmas propriedades da binomial negativa porém costuma se adequar melhor; Notação: `PIG(`μ, σ`)`; Função de ligação padrão: log para μ e log para σ.

Waring: também chamada de beta geométrica. Pode ser considerada uma geométrica com superdispersão; Notação: `WARING(`μ, σ`)`; Função de ligação padrão: log para μ e log para σ.

Logarítmica zero ajustada: é a distribuição logarítmica para casos com excesso de zeros. O parâmetro σ fornece a probabilidade exata da contagem ser nula; Notação: `ZALG(`μ, σ`)`; Função de ligação padrão: logit para μ e logit para σ.

Zipf zero ajustada: é a distribuição zipf para casos com excesso de zeros. O parâmetro σ fornece a probabilidade exata da contagem ser nula; Notação: `ZAZIPF(`μ, σ`)`; Função de ligação padrão: log para μ e logit para σ.

Poisson zero inflacionada: é a distribuição Poisson para casos com excesso de zeros. O parâmetro σ fornece a probabilidade exata da contagem ser nula; Notação: `ZIP`(μ, σ); Função de ligação padrão: log para μ e logit para σ.

- *Distribuições com três parâmetros:*

  Binomial beta negativa: adequada para dados com caudas pesadas; Notação: `BNB`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e log para ν.

  Delaporte: é resultado da convolução entre a Poisson e a binomial negativa; Notação: `DEL`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e logit para ν.

  Sichel: apropriada para dados com assimetria acentuada; Notação: `SICHEL`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e identidade para ν.

  Binomial negativa zero ajustada: é a binomial negativa para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade exata da contagem ser nula; Notação: `ZANBI`(μ, σ, ν) Função de ligação padrão: log para μ, log para σ e logit para ν.

  Poisson normal inversa zero ajustada: é a Poisson-normal inversa para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade exata da contagem ser nula; Notação: `ZAPIG`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e logit para ν.

  Binomial negativa zero inflacionada: é a binomial negativa para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade extra da contagem ser nula; Notação: `ZINBI`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e logit para ν.

  Poisson normal inversa zero inflacionada: é a Poisson-normal inversa para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade extra da contagem ser nula; Notação: `ZIPIG`(μ, σ, ν); Função de ligação padrão: log para μ, log para σ e logit para ν.

- *Distribuições com quatro parâmetros:*

  Poisson normal inversa deslocada generalizada: modela contagens com frequência com assimetria e/ou excesso de curtose; Notação: `PSGIG`(μ, σ, ν, τ); Função de ligação padrão: log para μ, log para σ, identidade para ν e logit para τ.

  Beta binomial negativa zero ajustada: é a beta binomial negativa para dados com excesso de zeros. τ é o parâmetro que fornece a probabilidade exata da contagem ser nula; Notação: `ZABNB`(μ, σ, ν, τ); Função de ligação padrão: log para μ, log para σ, log para ν e logit para τ.

  Sichel zero ajustada: é a Sichel para dados com excesso de zeros. τ é o parâmetro que fornece a probabilidade exata da contagem ser nula; Notação: `ZASICHEL`(μ, σ, ν, τ); Função de ligação padrão: log para μ, log para σ, identidade para ν e logit para τ.

  Beta binomial negativa zero inflacionada: é a beta binomial negativa para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade extra da contagem ser nula; Notação: `ZIBNB`(μ, σ, ν, τ); Função de ligação padrão: log para μ, log para σ, log para ν e logit para τ.

  Sichel zero inflacionada: é a Sichel para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade extra da contagem ser nula; Notação: `ZISICHEL`(μ, σ, ν, τ); Função de ligação padrão: log para μ, log para σ, identidade para ν e logit para τ.

Distribuições discretas binárias

- *Distribuições com um parâmetro:*

  Binomial: utilizada no estudo da probabilidade de sucesso de um evento; Notação: `BI`(n, μ); Função de ligação padrão: logit para μ.

- *Distribuições com dois parâmetros:*

  Beta binomial: é uma distribuição binomial em que a probabilidade do sucesso segue uma distribuição beta. Notação: `BB`(n, μ, σ); Função de ligação padrão: logit para μ e log para σ.

Binomial zero ajustada: é a binomial para dados com excesso de zeros. σ é o parâmetro que fornece a probabilidade exata da contagem ser nula; Notação: `ZABI`(n, μ, σ); Função de ligação padrão: logit para μ e logit para σ.

Binomial zero inflacionada: é a binomial para dados com excesso de zeros. σ é o parâmetro que fornece a probabilidade extra da contagem ser nula; Notação: `ZIBI`(n, μ, σ); Função de ligação padrão: logit para μ e logit para σ.

- *Distribuições com três parâmetros:*

Beta binomial zero ajustada: é a beta binomial para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade exata da contagem ser nula; Notação: `ZABB`(n, μ, σ, ν); Função de ligação padrão: logit para μ, log para σ, logit para ν.

Beta binomial zero inflacionada: é a beta binomial para dados com excesso de zeros. ν é o parâmetro que fornece a probabilidade extra da contagem ser nula; Notação: `ZIBB`(n, μ, σ, ν); Função de ligação padrão: logit para μ, log para σ, logit para ν.