UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
CURSO DE PÓS-GRADUAÇÃO EM ESPECIALIZAÇÃO EM ESTATÍSTICA E MODELAGEM QUANTITATIVA

MODELOS DE REGRESSÃO PARA O VALOR DA PRODUÇÃO DE ERVA-MATE NO RIO GRANDE DO SUL

MONOGRAFIA DE ESPECIALIZAÇÃO

Tatiane Fontana Ribeiro

Santa Maria, RS, Brasil

2019

MODELOS DE REGRESSÃO PARA O VALOR DA PRODUÇÃO DE ERVA-MATE NO RIO GRANDE DO SUL

Tatiane Fontana Ribeiro

Monografia apresentada ao Curso de Pós-Graduação em Especialização em Estatística e Modelagem Quantitativa da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do grau de

Especialista em Estatística e Modelagem Quantitativa

Orientador: Prof. Dr. Enio Júnior Seidel

Santa Maria, RS, Brasil

2019

Fontana Ribeiro, Tatiane

Modelos de regressão para o valor da produção de erva-mate no Rio Grande do Sul / por Tatiane Fontana Ribeiro. – 2019.
71 f.: il.; 30 cm.

Orientador: Enio Júnior Seidel
Monografia (Especialização) - Universidade Federal de Santa Maria, Centro de Ciências Naturais e Exatas, Curso de Pós-Graduação em Especialização em Estatística e Modelagem Quantitativa, RS, 2019.

1. Modelos lineares. 2. Modelos lineares generalizados. 3. Modelos aditivos generalizados para locação, escala e forma. 4. Erva-mate. 5. Rio Grande do Sul. I. Júnior Seidel, Enio. II. Título.

E-mail: tatianefontanaribeiro@gmail.com

Universidade Federal de Santa Maria
Centro de Ciências Naturais e Exatas
Curso de Pós-Graduação em Especialização em Estatística e Modelagem Quantitativa

A Comissão Examinadora, abaixo assinada,
aprova a Monografia de Especialização

MODELOS DE REGRESSÃO PARA O VALOR DA PRODUÇÃO DE ERVA-MATE NO RIO GRANDE DO SUL

elaborada por
Tatiane Fontana Ribeiro

como requisito parcial para obtenção do grau de
Especialista em Estatística e Modelagem Quantitativa

COMISSÃO EXAMINADORA:

Enio Júnior Seidel, Dr.
(Presidente/Orientador)

Augusto Maciel da Silva, Dr. (DE-UFSM)

Renata Rojas Guerra, Drª. (DE-UFSM)

Santa Maria, 19 de fevereiro de 2019.

RESUMO

Monografia de Especialização
Curso de Pós-Graduação em Especialização em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria

MODELOS DE REGRESSÃO PARA O VALOR DA PRODUÇÃO DE ERVA-MATE NO RIO GRANDE DO SUL

AUTORA: TATIANE FONTANA RIBEIRO
ORIENTADOR: ENIO JÚNIOR SEIDEL
Local da Defesa e Data: Santa Maria, 19 de fevereiro de 2019.

Modelos de regressão linear supondo a normalidade da variável resposta foram amplamente utilizados até novas técnicas de modelagem estatística mais flexíveis serem introduzidas. As situações reais são extremamente complexas, sendo que dificilmente dispõe-se de variáveis com características que satisfaçam ao modelo linear clássico. Devido a isso, foram propostas classes alternativas, dentre as quais citam-se os modelos lineares generalizados (MLG) e os modelos aditivos generalizados para locação, escala e forma (GAMLSS). A vantagem do MLG em relação ao modelo linear (ML) é que a variável resposta pode seguir qualquer distribuição de probabilidade pertencente à família exponencial. Na classe dos modelos GAMLSS há maior flexibilidade em relação à distribuição da variável resposta, cujo modelo probabilístico não precisa, necessariamente, pertencer à família exponencial de distribuições. Ademais, é possível modelar outros parâmetros da distribuição em termos das covariáveis. Considerando a relevância da cultura permanente da erva-mate para a economia do Rio Grande do Sul, a quantidade de estatísticas públicas disponíveis acerca do tema e as técnicas de modelagem estatística mencionadas, objetiva-se obter um modelo de regressão que explique a variação do valor da produção deste produto agrícola no estado gaúcho. O conjunto de dados considerado é disponibilizado pela Fundação de Economia e Estatística. Este contém a variável resposta: valor da produção de erva-mate e covariáveis quantitativas relacionadas a sua produção e comercialização em 2016. São acrescentadas a este banco de dados as covariáveis qualitativas referentes aos polos ervateiros gaúchos e microrregiões tratadas como variáveis *dummy*. A partir disso, realiza-se uma estatística descritiva com intuito de identificar o comportamento de todas as variáveis. São ajustados modelos de regressão de acordo com cada classe mencionada, desde a técnica clássica (ML) até a mais sofisticada (GAMLSS). Evidencia-se a superioridade e boa qualidade do ajuste dos modelos GAMLSS em relação as demais classes de acordo com os critérios de seleção de modelos considerados e as análises gráficas dos resíduos. No modelo GAMLSS final, entre as covariáveis quantitativas, foram significativas ao nível de 5% a quantidade produzida, a área colhida e a área destinada à colheita com contribuição positiva à variável resposta, corroborando com a relação linear positiva evidenciada na análise de correlação. O polo Planalto Missões foi o único significativo e apresentou efeito positivo ao valor de produção, uma vez que é o segundo maior prodrutor estadual. Foram significativas as microrregiões: Caxias do Sul, Erechim, Frederico Westphalen, Gramado-Canela, Guaporé, Lajeado-Estrela, Santa Cruz do Sul, Santa Rosa e Três Passos. Com exceção de Gramado-Canela, todas apresentaram efeito negativo à variável dependente.

Palavras-chave: Modelos lineares. Modelos lineares generalizados. Modelos aditivos generalizados para locação, escala e forma. Erva-mate. Rio Grande do Sul.

ABSTRACT

Specialization Monograph
Specialization Course in Statistics and Quantitative Modelling
Federal University of Santa Maria

REGRESSION MODELS TO YERBA MATE PRODUCTION VALUE IN RIO GRANDE DO SUL

AUTHOR: TATIANE FONTANA RIBEIRO
ADVISOR: ENIO JÚNIOR SEIDEL
Defense Place and Date: Santa Maria, February 19st, 2019.

Linear regression models that suppose normal response variables were widely used until more flexible techniques were introduced. Real situations are very complex, thus it is difficult that variables comply with the assumptions of the classic linear model, like the normality of the response variable. Alternative techniques were then proposed, like: the generalized linear models (GLM) and generalized additive models for location, scale and shape (GAMLSS). GLM main advantage in relation to linear model (LM) technique, because the response variable can follow any distribution of the exponential family, besides of the normal distribution. It is also possible to model other parameters of the distribution based on the covariables. By considering the relevance of the permanent culture of yerba mate to the economy RS, the quantity of public statistics averable about the subject and the techniques of modelling statistic mentioned, the of purpose of this study is to obtain a regression model to explain the variation of the production value of the yerba mate in the RS. The data set was obtained from Foundation of Economy and Statistics. This data set has the response variable: production value of yerba mate and quantitative covariables associated to its production and commercialization in 2016. It is added to data set qualitative covariables referring to poles and microregions studied like dummy variables. Thereafter, descriptive statistic analysis is done in order to identify the behavior of all variables. Regression models are obtained from classical technique (LM) to the more sophisticated (GAMLSS). It is noted that the GAMLSS model had the best fit according with models selection criteria and graphical analysis of the residuals. Final GAMLSS model at the 5% significance level the significant quantitative covariables were: quantity produced, harvested area and area for harvesting. This covariables present positive contribution to response variable according linear relation showed through the correlation analysis. Polo Planalto Missões was the only significant and it presented positive effect to production value, because it is second largest state producer. The significant microrregions were: Caxias do Sul, Erechim, Frederico Westphalen, Gramado-Canela, Guaporé, Lajeado-Estrela, Santa Cruz do Sul, Santa Rosa e Três Passos. Excepting Gramado-Canela microrregion, the others microrregions presented negative effect to dependent variable.

Keywords: Linear models. Generalized linear models. Generalized additive models for location, scale and shape. Yerba mate. Rio Grande do Sul.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Por muito tempo o estudo da associação entre variáveis por meio de técnicas de regressão restringiu-se à utilização dos modelos de regressão lineares (ML) propostos por Galton (GALTON, 1886). Nesta classe, é necessário que sejam atendidos alguns pressupostos, entre os quais destaca-se a condição de que a variável resposta seja normalmente distribuída. Contudo, as situações reais são extremamente complexas, sendo que dificilmente dispõe-se de variáveis com esta característica. Foram propostas então, alternativas de transformação dos dados para atender a suposição de normalidade da variável dependente (BARTLETT, 1947; BOX; COX, 1964; ANDREWS, 1971; CARROLL, 1980; HERNANDEZ; JOHNSON, 1980; BICKEL; DOKSUM, 1981).

Há casos, porém, em que não é possível ajustar modelos lineares mesmo transformando a variável resposta. Para suprir esta e outras limitações dos ML, Nelder e Wedderburn (1972) propuseram a classe dos modelos lineares generalizados (MLG). A vantagem desta em relação à anterior é que a variável resposta não precisa seguir uma distribuição normal, mas sim qualquer modelo probabilístico que pertença à família exponencial de distribuições.

Também foram propostos os modelos aditivos generalizados (MAG) por Hastie e Tibshirani (1990). Embora mais flexíveis que os ML não possibilitam modelar variável resposta cuja distribuição não faça parte da família exponencial, bem como consideram que o parâmetro de dispersão não varia de acordo com as covariáveis. Extensões para os MLG e MAG com base em diferentes métodos de estimação visando possibilitar a inclusão de termos aleatórios são dadas por McCulloch (1997) e Lin e Zhang (1999).

Devido a crescente demanda de análise estatística de uma grande quantidade de dados e necessidade de obter modelos mais realísticos para descrever situações complexas, Ribgy e Stasinopoulos (2005) propuseram a classe dos modelos aditivos generalizados para posição, escala e forma (GAMLSS). Os modelos GAMLSS são fortemente flexíves na especificação da distribuição da variável resposta, que pode extrapolar os modelos pertencentes à família exponencial. Possibilitam também que a distribuição de todos os parâmetros do modelo probabilístico da variável resposta seja modelada como função das covariáveis. Além disso, o modelo pode conter termos paramétricos, não paramétricos e aleatórios, o que fornece flexibidade extra à classe (STASINOPOULOS; RIGBY, 2007). Por isso, possuem como casos especiais as demais classes já mencionadas.

No cenário de análise estatística de dados referentes à produção e comercialização de erva-mate no estado do Rio Grande do Sul (RS) são encontrados na literatura apenas trabalhos de experimentação agrícola, análises de estatísticas descritivas e ajustes de ML clássicos (STORCK et al., 2002; BERGER, 2007; PICOLOTTO et al., 2013; RIGO et al., 2014; OLIVEIRA; WAQUIL, 2015; CHECHI; SCHULTZ, 2016; VOGT; NEPPEL; SOUZA, 2016; ZANIN; MEYER, 2018).

A erva-mate é uma atividade não madereira que compõe o merdado agroflorestal brasileiro (OLIVEIRA; WAQUIL, 2015). Esta cultura possui grande importância econômica, principalmente no RS, bem como destaca-se no cenário nacional devido a contribuições no processo de desenvolvimento regional, por meio das esferas econômica, social e ambiental (PICOLOTTO et al., 2013). A tradição do consumo do chimarrão, por exemplo, impulsiona tanto a produção quanto a comercializaçao da erva-mate do estado gaúcho (OLIVEIRA; WAQUIL, 2015). O RS lidera em termos de quantidade produzida média de 2014 à 2016, contando com uma produção de 288.586 toneladas, comparando-o com os demais estados produtores a saber: Paraná (PR), Santa Catarina (SC) e Mato Grosso do Sul (MS) (IBGE, 2016).

A partir de 1995/1996 a erva-mate foi incluída pelo IBGE entre as nove culturas mais importantes do RS (IBGE, 1996). As informações são disponibilizadas pela pesquisa Produção Agrícola Municipal (PAM). Salienta-se que o conjunto de produtos das lavouras temporárias e permanentes nacionais investigados é selecionado tanto pela extrema relevância econômica na pauta de exportações, como também por sua relevância social, sendo o município a unidade de coleta. Assim, dispõe-se de informações referentes à área plantada, área destinada à colheita, área colhida, quantidade produzida, rendimento médio e preço médio pago ao produtor de erva-mate (IBGE, 2017).

Dada a relevância econômica e social da cultura permanente de erva-mate, a quantidade de estatísticas públicas e ferramentas de análise estatística disponíveis, objetiva-se com este estudo determinar um modelo de regressão para o valor da produção de erva-mate no RS, considerando dados de 2016. Prentende-se predizer este valor a partir de uma combinação de covariáveis e identificar quais destas e de que forma influenciam na sua variação. Para tanto, ajustam-se modelos de regressão para esta variável resposta considerando as três técnicas de modelagem estatísticas fortemente discutidas na literatura: ML, MLG e GAMLSS.

1.1 Objetivos

1.1.1 Objetivo geral

Determinar um modelo de regressão para o valor da produção de erva-mate em folha verde no estado do RS no ano de 2016, a fim de explicar seu valor a partir de uma combinação de covariáveis e identificar quais destas e de que forma influenciam na sua variação.

1.1.2 Objetivos específicos

• Realizar uma análise descritiva dos dados de produção da erva-mate em folha verde nos municípios gaúchos.

• Identificar covariáveis que possam explicar a variabilidade no valor da produção de erva-mate.

• Entender as relações entre as variáveis geradoras do valor de produção de erva-mate em folha verde com base no modelo determinado.

1.2 Justificativa

A obtenção de modelos de regressão capazes de explicar um fenômeno consiste em uma das principais técnicas estatísticas que possibilitam obter uma versão simplificada de algum problema ou de determinada situação real (EMILIANO, 2013). Assim, um modelo que se ajusta adequadamente ao conjunto de dados pode auxiliar na tomada de decisão do tema/problema em estudo.

Na busca pelo modelo que capte a essência do problema em pauta, de modo que confirme também a relevância lógica e teórica das covariáveis associadas à variável resposta é necessário considerar dois importantes fatos. O primeiro é que aumentar o número de covariáveis eleva o grau de complexidade na interpretação do modelo ajustado. E o segundo é que modelar um fenômeno com base em poucas covariáveis, embora proporcione uma interpretação mais simples, pode levar a um ajuste não adequado que forneça conclusões distorcidas da realidade (FLORENCIO, 2010).

O ideal é determinar um modelo intermediário entre o modelo com o mínimo e o máximo de covariáveis (FLORENCIO, 2010). Escolher o melhor modelo corresponde à atingir um equilíbrio entre um bom ajustamento ao conjunto de dados, um modelo parcimonioso e in-

terpretável (TURKMAN; SILVA, 2000). Neste sentido, justifica-se buscar o ajuste de modelos considerando desde as classes de modelagem estatísticas mais simples às mais complexas que se dispõe. Por isso, são estudadas na presente monografia desde a técnica clássica de regressão tradicional (ML) até a classe de modelos GAMLSS, no ajuste de modelos de regressão para o valor da produção de erva-mate no RS.

Ademais justifica-se também o estudo devido a relevância da produção e comercilaização da erva-mate para a economia do estado do RS e a pequena quantidade de trabalhos acadêmicos pertinentes ao tema. Assim, verifica-se a importância de estudar o comportamento do valor da produção da erva-mate de acordo com covariáveis que possam explicar sua variação.

1.3 Estrutura do trabalho

O trabalho está organizado como segue. No capítulo 2 é apresentado o contexto histórico dos modelos de regressão, assim como definições e formas de estimação dos parâmetros acerca das três classes supracitadas: ML, MLG e GAMLSS. Também mencionam-se quais critérios de seleção dos modelos foram utilizados. Ainda neste capítulo, aborda-se um breve referencial teórico sobre a cultura permanente de erva-mate, com ênfase em estatísticas que ilustram sua importância ao cenário econômico do estado gaúcho. O capítulo 3 apresenta os materiais e métodos utilizados, em que descreve-se o conjunto de dados e as covariáveis que são acrescentadas, bem como a metodologia para obtenção das análises estatísticas e modelos de regressão. No capítulo 4 apresenta-se a estatística descritiva do banco de dados considerado, seguida dos modelos de regressão ajustados de acordo com cada uma das três técnicas e comparações entre estes. E, no capítulo 5 são apresentadas as considerações finais do estudo.

2 REFERENCIAL TEÓRICO

Neste capítulo são contextualizados historicamente os modelos de regressão de uma forma geral. Discutem-se com base na literatura as classes de modelos: ML, MLG e GAMLSS, bem como apresenta-se a forma com que são estimados os parâmetros do modelo em cada uma. Além disso, aborda-se uma breve fundamentação teórica acerca da cultura permanente de erva-mate e sua relevância na economia do estado RS.

2.1 Contexto histórico dos modelos de regressão

No início do século XIX, Legendre e Gauss desenvolveram o modelo de regressão linear clássico. O modelo proposto consistiu na principal técnica de modelagem estatística até meados do século seguinte. Contudo, outros modelos não lineares ou que não atendiam ao pressuposto de normalidade dos dados foram introduzidos para possibilitar a modelagem de situações nas quais o modelo linear não era adequado (TURKMAN; SILVA, 2000).

A origem da análise de regressão é atribuída à Francis Galton, um antropólogo, meteorologista, matemático e estatístico inglês (1822-1911). Com o objetivo de expressar por meio de fórmulas matemáticas a relação entre a altura de um homem específico e seus filhos, elaborou um artigo no qual definiu a regressão em direção a média. Galton (1886) constatou que existia uma tendência de que pais altos tivessem filhos altos, assim como pais baixos tivessem filhos baixos. Entretanto, verificou que a altura de crianças nascidas de pais mais altos ou mais baixos tendia a atingir a altura média da população como um todo, regredindo ou aumentando, respectivamente.

Esta tendência foi confirmada, posteriormente, nos estudos de Karl Pearson. Após coletar um banco de dados contendo mais de mil alturas de membros de grupos familiares, Pearson (1903) evidenciou que filhos de um grupo de pais altos possuíam altura média menor que a de seus pais e que filhos de um grupo de pais baixos tinham altura média maior que a de seus pais. Em síntese, de fato, a altura média dos filhos tendia à altura média de toda a população.

Mais tarde, foram introduzidos modelos de regressão para variáveis respostas discretas. Fisher (1922) propôs o modelo binomial com ligação complementar log-log para ensaios de diluição, que consiste em uma solução completa para avaliar este tipo de ensaio. Bliss (1935) introduziu o modelo probit para modelar dados no intervalo unitário, definindo o probit como

um método de estimação de doses críticas em ensaios de dose-resposta. Berkson (1944) introduziu o modelo logit, ficando amplamente reconhecido como o principal proponente do uso da regressão logística em detrimento de modelos lineares, uma vez que ajustou um modelo logístico consideravelmente mais simples que o modelo normal para probits defendido por Fisher (1922) e Bliss (1935).

Dados binários com preditores categóricos com base na regressão logística foram modelados por Dyke e Patterson (1952). Os autores demonstraram que regressão logística com um modelo de ANOVA facilita a compreensão e interpretação do modelo ajustado. Rasch (1960) propôs um modelo teórico de resposta a itens para itens dicotômicos, no qual a variável resposta é modelada como função de dois parâmetros: pessoa e item.

Posteriormente, Birch (1963) desenvolveu os modelos log-lineares para ajustar dados de contagens envolvendo a distribuição de Poisson. O autor evidenciou que em uma tabela de contigência os totais marginais observados, as estimativas de máxima verossimilhança dos parâmetros do modelo e os totais marginais do conjunto de frequências ajustadas com base nestas estimativas são proporcionais e fornecem, portanto, a mesma estimativa para os parâmetros comuns do modelo log-linear.

No ramo de análise de sobrevivência, Feigl e Zelen (1965) introduziram o modelo exponencial, propondo um método de estimar distribuições de sobrevivência quando o tempo de sobrevivência segue distribuições exponencias simples, com um parâmetro diferente para cada paciente. O modelo estatístico foi generalizado para permitir a estimativa de máxima verossimilhança dos parâmetros da regressão linear por Zippin e Armitage (1966).

Glasser (1967) propôs a regressão exponencial, por meio do estudo de um modelo exponencial de riscos proporcionais para comparar a experiência de mortalidade de dois ou mais grupos de indivíduos considerando que a força de mortalidade para cada indivíduo é uma função da covariável.

Modelos com polinômios inversos envolvendo a distribuição Gama com função de ligação recíproca para experimentos, nos quais um ou mais fatores quantitativos são testados, cada um contendo dois ou mais niveis foram apresentados por Nelder (1966).

A partir deste contexto histórico, apresentam-se três técnicas de regressão que são bastante discutidas e utilizadas na literura atualmente: ML, MLG e GAMLSS.

2.2 Modelos lineares

A análise de regressão linear é, frequentemente, utilizada na resolução de inúmeros problemas de diferentes campos da Ciência. Tem como objetivo estudar a relação linear entre, no mínimo, duas variáveis quantitativas, isto é, entre uma variável dependente e uma ou mais variáveis independentes.

O ML descreve a variável resposta (Y) como soma de uma combinação linear de parâmetros desconhecidos com covariáveis independentes e um erro aleatório. Também é possível expressar o valor esperado de Y como uma função de polinômio de maior grau de uma única covariável. O vetor aleatório erro ($\boldsymbol{\epsilon}$) segue uma distribuição normal $N(0, \sigma^2)$, fazendo com que a variável resposta também seja distribuída normalmente (CHARNET et al., 2008).

Dado que algumas suposições básicas para o modelo obtido sejam satisfeitas, as técnicas de regressão possibilitam determinar equações que permitem realizar boas estimativas, precisas e eficientes. A avaliação da equação ajustada é feita mediante a aplicação de inúmeros testes estatísticos que permitem inferir conclusões à população e ajustar a melhor equação possível (SCHNEIDER; SCHNEIDER; SOUZA, 2009).

2.2.1 Definição

Seja o vetor de observações $\mathbf{y}$ de dimensão n correspondente a uma realização da variável aleatória $\mathbf{Y}$, distribuída independentemente com média $\boldsymbol{\mu}$, e p parâmetros, $\beta_1, \cdots, \beta_p$, o modelo linear ordinário é definido matricialmente por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

e

$$\boldsymbol{\epsilon} \sim NM_n(0, \sigma^2\mathbf{I}),$$

com

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{e} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

em que os componentes do vetor $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}^T$ correspondem às n observações da variável resposta. A distribuição de $\mathbf{y}$ é normal multivariada com vetor de médias $\mathbf{X}\boldsymbol{\beta}$ e matriz

de variância $\sigma^2\mathbf{I}$, isto é, $\mathbf{y} \sim N_m\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$, resultado que decorre da distribuição do erro aleatório. A matriz $\mathbf{X}$ de dimensão $n \times (p+1)$ corresponde ao conjunto de variáveis explicativas, sendo que cada linha é referente a uma observação diferente e cada coluna a uma variável explicativa. O vetor $\boldsymbol{\beta}$ de dimensão $(p+1)$ expressa o conjunto dos parâmetros desconhecidos, denominados de coeficientes de regressão (CHARNET et al., 2008).

2.2.2 Estimação dos parâmetros

A estimativa dos parâmetros do ML, geralmente é feita através da utilização do método dos mínimos quadrados ordinários ou mínimos quadrados generalizados (CHARNET et al., 2008; GUJARATI; PORTER, 2011; LINDSEY, 1997). Nenhum desses métodos exige que a distribuição de $\mathbf{y}$ seja conhecida (MCCULLOCH; SEARLE, 2001).

Um método de estimação pontual que possui propriedades teóricas mais fortes que as do método de mínimos quadrados ordinários é o método de estimação por máxima verossimilhança. Contudo, desde que a distribuição do erro seja normal no ML, os estimadores de mínimos quadrados são equivalentes aos estimadores de máxima verossimilhança (EMV) (GUJARATI; PORTER, 2011). Deste modo, apresentam-se os estimadores do ML, obtidos via aplicação do método de máxima verossimilhança, uma vez que possui ótimas propriedades, tais como, consistência e eficiência assintótica (DEMÉTRIO; CORDEIRO, 2007).

A partir de (2.1), escreve-se a função de verossimilhança, dada por

$$L\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}\right) = \frac{\exp[-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{I}/\sigma^2\right)\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)]}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}N} \tag{2.2}$$

de (2.2) segue que a log-verossimilhança é dada por

$$\ell\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}\right) = -\frac{1}{2}N\log\left(2\pi\right) - \frac{1}{2}N\log\sigma^2 - \frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)/\sigma^2. \tag{2.3}$$

Derivando a equação (2.3) em relação aos parâmetros $\boldsymbol{\beta}$ e σ^2, igualando as derivadas parciais a zero e substituindo $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$ tem-se que

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

e

$$\hat{\sigma}^2 = \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)/N,$$

se $\mathbf{X}'\mathbf{X}$ existir e se a dimensão da matriz do modelo ($\mathbf{X}$) possui dimensão $N \times p$ possuir posto de coluna completo p. Esta última condição é muito restritiva, pois são inúmeras as situações em que isso não acontece (MCCULLOCH; SEARLE, 2001).

2.3 Modelos lineares generalizados

A classe dos Modelos Lineares Generalizados (MLG) é uma extensão dos modelos lineares normais, definida por Nelder e Wedderburn (1972), que recebe destaque na literatura. O avanço na proposta destes modelos é que a variável resposta (dependente) pode assumir outras distribuições de probabilidade além da distribuição normal. Diferentemente do ML, no qual a média é tomada como uma combinação linear de parâmetros, no MLG, a relação entre as variáveis explanatórias e a média da variável resposta é estabelecida por meio de uma função de ligação que não precisa ser, necessariamente, a função identidade (MCCULLOCH; SEARLE, 2001).

Embora esta classe de modelos também exija independência das observações, problemas como falta de normalidade e homocedasticidade, frequentes em MLs, não ocorrem. O que justifica este fato é que a variância é dada como uma função da média e a aditividade dos efeitos decorre naturalmente como propriedade das respostas esperadas (MCCULLAGH; NELDER, 1989).

Na seleção do modelo, o principal problema é escolher as covariáveis a serem consideradas na estrutura de regressão. Segundo Cordeiro e Demétrio (2007) os MLGs são constituídos por uma variável dependente que possui distribuição pertencente à família de distribuições exponencial, variáveis independentes e uma função de ligação que descreve a relação funcional entre a variável resposta e as variáveis independentes.

Devido ao leque de distribuições às quais pode pertencer a variável dependente, os MLGs possibilitam a modelagem de variáveis que assumem a forma discreta, contínua, simétrica e assimétrica. Embora haja, portano, maior flexibilidade na modelagem estatística por meio desta técnica unificadora, há a limitação de que os erros sejam independentes, dificultanto, por exemplo, a modelagem de banco de dados com estruturas longitudinais.

2.3.1 Definição

O ML definido em (2.1) pode ser escrito de forma equivalente a

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu},$$

onde

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{2.4}$$

Seguindo as três especificações que constituem o trinômio necesário à definição de um MLG, é possível generalizar (2.4) da seguinte forma

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{2.5}$$

em que $\mathbf{Y}$ é a variável aleatória dependente, $\mathbf{X}$ é a matriz de covariáveis do modelo associada aos parâmetros por meio do preditor linear $\boldsymbol{\eta}$, que relaciona-se à $\mathbf{Y}$ pela função de ligação $g(\cdot)$, monótona e diferenciável.

Nesta formulação a variável dependente pode assumir qualquer distribuição da família exponencial e a função de ligação é equivalente a qualquer função diferenciável e monótona. Além disso, se $\mathbf{Y}$ possui distribuição normal e $g(\cdot)$ é igual a função identidade, então (2.5) é equivalente a (2.2.1) (MCCULLAGH; NELDER, 1989).

Antes de apresentar a estimação dos parâmetros do MLG, seguem nas próximas subseções alguns conceitos referentes à distribuição da variável resposta nesta classe, bem como maiores detalhes sobre funções de ligação.

2.3.2 Família exponencial

Na teoria dos MLG a família exponencial é muito importante por permitir que sejam incorporados dados assimétricos, de natureza discreta ou contínua e duplamente limitados (DEMÉTRIO; CORDEIRO, 2007). Uma variável aleatória $\mathbf{Y}$ tem distribuição da família exponencial se sua função densidade ou de probabilidade é

$$f_Y(y;\theta,\phi) = \exp\{(y\,\theta - b(\theta))/a(\phi) + c(y,\phi)\} \tag{2.6}$$

para funções específicas $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$. Caso ϕ seja conhecido, (2.6) é modelo da família exponencial com parâmetro canônico θ (MCCULLAGH; NELDER, 1989). De (2.6), escrevendo

a log-verossimilhança, tem-se que

$$\ell(\theta; y) = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi). \tag{2.7}$$

Considerando as relações conhecidas da função escore em que

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \tag{2.8}$$

e

$$E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + E\left(\frac{\partial \ell}{\partial \theta}\right)^2 = 0, \tag{2.9}$$

de (2.7) e (2.8) tem-se que

$$E(Y) = \mu = b'(\theta). \tag{2.10}$$

Analogamente segue de (2.7), (2.8) e (2.9) que

$$var(Y) = b''(\theta)\, a(\phi). \tag{2.11}$$

A variância de $\mathbf{Y}$ é produto de duas funções, sendo que uma delas é função apenas do parâmetro canônico, consequentemente da média e a outra apenas do parâmetro de dispersão ϕ. Comumente a função $a(\phi)$ é definida como o quociente de ϕ, também denotado por σ^2 constante em todas as observações, pelo peso ω que varia a cada observação (MCCULLAGH; NELDER, 1989).

Nos MLGs a variância, assimetria e curtose são dadas de forma implícita por meio de sua dependência do parâmetro de locação (μ) e não explicitamente em termos das variáveis independentes (RIGBY; STASINOPOULOS, 2005).

2.3.3 Função de ligação

A relação entre o preditor linear η e o valor esperado de um dado y é estabelecida por meio de uma função de ligação. De acordo com a distribuição da variável independente são estabelecidas as funções de ligação correspondentes. Quando o parâmetro canônico θ é igual ao preditor linear η tem-se uma função de ligação canônica (MCCULLAGH; NELDER, 1989). Isso garante, que se a variável resposta segue uma distribuição de Poisson em que $\mu > 0$, por exemplo, não sejam estimados valores fora dos reais positivos para a média já que η pode ser negativo. Para este caso, utiliza-se, geralmente a função de ligação canônica $g(\mu) = \log(\mu)$, cuja inversa é $\mu = \mathrm{e}^{\eta}$

Tabela 2.1 – Funções de ligação canônicas na família exponencial

Distribuição	**Função de Ligação**
Normal	$\eta = \mu$
Poisson	$\eta = \log \mu$
Binomial	$\eta = \log\{\pi/(1-\pi)\}$
Gama	$\eta = \mu^{-1}$
Gaussiana inversa	$\eta = \mu^{-2}$
Exponencial	$\eta = \log \mu$

Fonte: Adaptado de Demétrio e Cordeiro (2007).

Na Tabela 2.1 são apresentadas as funções $g(\cdot)$ canônicas para algumas distribuições da família exponencial.

Os modelos decorrentes são denominados canônicos. Frequentemente a escolha de funções de ligação canônicas resulta em uma escala adequada para a modelagem com interpretação prática para os parâmetros de regressão. Ademais se tem outras vantagens teóricas, como a existência de um conjunto de estatísticas suficientes para $\boldsymbol{\beta}$ e simplificações no algoritmo de estimação (DEMÉTRIO; CORDEIRO, 2007).

Desse modo, o problema na modelagem via aplicação de um MLG consiste na escolha do trinômio fundamental: distribuição da variável dependente, matriz modelo e função de ligação. De acordo com Cordeiro e Demétrio (2007), esta decisão pode ser tomada mediante análise dos dados ou a partir de alguma experiência anterior do pesquisador.

2.3.4 Estimação dos parâmetros

De forma similar aos MLs, os parâmetros $\beta's$ podem ser estimados por vários métodos. Contudo, devido à propriedades como consistência e eficiência assintótica dos EMVs, será utilizado o método de máxima verossilmilhança (DEMÉTRIO; CORDEIRO, 2007).

O logaritmo da função de verossimilhança, considerando-se o parâmetro de dispersão, ϕ conhecido dado o vetor $\mathbf{y}$ e usando-se a expressão (2.7), é dado por

$$\ell(\beta) = \frac{1}{\phi}\sum_{i=1}^{n}\left[y_i\theta_i - b(\theta_i)\right] + \sum_{i=1}^{n} c(y_i, \phi), \tag{2.12}$$

em que $\theta_i = q(\mu_i)$, $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r$ (DEMÉTRIO; CORDEIRO, 2007).

Pela regra da cadeia, de (2.12), calcula-se o vetor escore $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ de dimensão p, com

$$U_r = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{i=i}^{n} \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_r} \tag{2.13}$$

e, das expressões (2.10) e (2.11) tem-se que

$$U_r = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)\,\frac{1}{V_i}\frac{d\mu_i}{d\eta_i}x_{ir}, \tag{2.14}$$

em que $r = 1, \cdots, p$. Como de forma usual, obtém-se o estimador $\hat{\boldsymbol{\beta}}$ do vetor de parâmetros $\boldsymbol{\beta}$ igualando-se as funções escores a zero, (2.13) e (2.14). Porém, geralmente estas funções não são lineares e são resolvidas numericamente por processos iterativos do tipo Newton-Raphson (DEMÉTRIO; CORDEIRO, 2007).

Aplicando o método iterativo de Newton-Raphson com base na aproximação de Taylor e realizando algumas manipulações e cálculos obtém-se a equação matricial

$$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{z}^{(m)}, \tag{2.15}$$

que é valida para qualquer MLG. A eq. (2.15) evidencia que para determinar a solução das equações de máxima verossimilhança, pode-se, equivalentemente, calcular de maneira repetida uma regressão linear ponderada de uma variável resposta ajustada $\mathbf{z}$ sobre a matriz $\mathbf{X}$ por meio da utilização de uma função peso $\mathbf{W}$ que se altera durante o processo de iteração. Assim, com o auxílio de algum software estatístico, implementando (2.15) obtém-se os estimadores de máxima verossimilhança dos parâmetros lineares do MLG (DEMÉTRIO; CORDEIRO, 2007).

2.4 Modelos aditivos generalizados de locação, escala e forma

A classe dos modelos aditivos generalizados que consideram posição, escala e forma (GAMLSS) foi proposta por Rigby e Stasinopoulos (2001) e Rigby e Stasinopoulos (2005) com intuito de superar as limitações dos clássicos MLGs e Modelos Aditivos Generalizados (MAG).

Nesta classe também assume-se que a variável resposta consista em observações independentes, porém esta pode pertencer a uma ampla família de distribuições (contínuas e discretas) que vai além da família exponencial e que acomoda elevadas assimetrias e/ou curtoses. Além disso, os modelos GAMLSS, permitem que sejam modelados outros parâmetros da distribuição da variável dependente e não apenas a média (ou locação) como nos modelos populares mencionados (STASINOPOULOS; RIGBY, 2007).

2.4.1 Definição

Seja o vetor de dimensão n, $\mathbf{y}^T = \{y_1, y_2, \ldots, y_n\}$ correspondente às observações da variável resposta. A função de ligação monótona e conhecida $g_k(\cdot)$ relaciona os parâmetros $\boldsymbol{\theta_k}$, para $k = 1, 2, \cdots, p$, às variáveis independentes e efeitos aleatórios através do modelo aditivo dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.16}$$

em que, ambos os vetores $\boldsymbol{\theta}_k^T$ e $\boldsymbol{\eta}_k$ possuem tamanho n, sendo $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \cdots, \theta_{nk})$ e $\boldsymbol{\beta}_k^T = \left(\beta_{1k}, \beta_{2k}, \cdots, \beta_{J_k'k}\right)$ o vetor de parâmetros. A matriz do modelo $\mathbf{X}_k$ é conhecida de ordem $n \times J_k'$, $\mathbf{Z}_{jk}$ é uma matriz conhecida fixa de ordem $n \times q_{jk}$ e γ_{jk} é uma variável aleatória $q_{jk}-$dimensional (RIGBY; STASINOPOULOS, 2005).

Se $J_k = 0$, para $k = 1, 2, \cdots, p$ a expressão (2.16) é reduzida a um modelo totalmente paramétrico definido por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k. \tag{2.17}$$

A formulação (2.16) possibilita que sejam incorporados diferentes tipos de efeitos aleatórios aditivos no modelo. Detalhes dessas combinações, bem como casos particulares podem ser encontrados em Rigby e Stasinopoulos (2005).

Usualmente, os dois primeiros parâmetros, $\boldsymbol{\theta_1}$ e $\boldsymbol{\theta_2}$ caracterizam a locação (μ) e a escala (σ). Para muitas famílias de distribuições, dois parâmetros de forma são suficientes: $\boldsymbol{\nu}\,(=\boldsymbol{\theta}_3)$ e $\boldsymbol{\tau}\,(=\boldsymbol{\theta}_4)$ (RIGBY; STASINOPOULOS, 2005). Assim, tem-se o modelo

$$\left.\begin{array}{l} g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\beta_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\gamma_{j1} \\ g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\beta_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\gamma_{j2} \\ g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\beta_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\gamma_{j3} \\ g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\beta_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\gamma_{j4} \end{array}\right\} \tag{2.18}$$

É evidente, assim, que os modelos GAMLSS apresentam maior flexibilidade, não somente em termos da distribuição da variável resposta que não precisa, necessariamente, pertencer à família exponencial, como a possibilidade de modelar todos os parâmetros da sua distribuição de probabilidade.

De modo geral, os parâmetros μ, σ, ν e τ representam a locação, escala, assimetria e curtose, respectivamente, embora possam ser quaisquer outros parâmetros da distribuição. O

pacote `gamlss` implementado em `R` permite escolher cerca de 100 distribuições com quatro parâmetros para a variável resposta (STASINOPOULOS; RIGBY, 2007).

Na Tabela 2.2 são apresentadas algumas das distribuições de probabilidade da família `gamlss` ($\mathcal{D}$) com suporte no intervalo $(0, \infty)$, bem como os respectivos comandos em `R` e as funções de ligação padrão da função `gamlss()` para todos os parâmetros de cada modelo.

Tabela 2.2 – Distribuições com suporte nos $\mathbb{R}_+^*$ implementadas no pacote *gamlss*

Distribuições	**Nomenclatura em** `R`	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\nu}$	$\boldsymbol{\tau}$
Box-Cox, Cole e Green	BCCG()	identity	log	identity	-
Box-Cox Cole-Green orig.	BCCGo()	log	log	identity	-
Exponencial de potência Box-Cox	BCPE()	identity	log	identity	log
Exponencial de potência orig.	BCPEo()	log	log	identity	log
Box-Cox-t	BCT()	identity	log	identity	log
Box-Cox t orig.	BCTo()	log	log	identity	log
Exponencial	EXP()	log	-	-	-
Gama	GA()	log	log	-	-
Beta generalizada tipe 2	GB2()	log	log	log	log
Gama Generalizada	GG()	log	log	identity	-
Gaussiana Inversa Generalizada	GIG()	log	log	identity	-
Gamma inversa	IGAMMA()	log	log	-	-
Gaussiana Inversa	IG()	log	log	-	-
Log-Normal	LOGNO()	log	log	-	-
Log-Normal 2	LOGNO2()	log	log	-	-
Log-Normal (Box-Cox)	LNO()	log	log	fixed	-
Pareto 2	PARETO2()	log	log	-	-
Pareto 2 original	PARETO2o()	log	log	-	-
Pareto 2 repar.	GP()	log	log	-	-
Weibull	WEI()	log	log	-	-
Weibull (PH)	WEI2()	log	log	-	-
Weibull (μ the mean)	WEI3()	log	log	-	-

Fonte: Adaptado de Rigby, Stasinopoulos e Bastiani (2017).

2.4.2 Estimação dos parâmetros

Na estrutura dos modelos GAMLSS é suposto um efeito aleatório com distribuição normal no preditor linear. O resultado da estimação utiliza matriz de suavização mediante um algoritmo de retroajuste (*backfitting*). Assim, assume-se no modelo (2.16) que γ_{jk}, à priori, possui distribuição normal independente com $\gamma_{jk} \sim N_{q_{jk}}\left(\mathbf{0}, \mathbf{G}_{jk}^{-}\right)$, sendo que $\mathbf{G}_{jk}^{-}$ é a inversa generalizada de uma matriz simétrica de ordem $q_{jk} \times q_{jk}$, $\mathbf{G}_{jk} = \mathbf{G}_{jk}\left(\boldsymbol{\lambda}_{jk}\right)$. A matriz $\mathbf{G}_{jk}$ pode depender de um vetor de hiper-parâmtros $\boldsymbol{\lambda_{jk}}$ e caso seja singular γ_{jk} é dita ter uma função densidade proporcional a $exp\left(-\frac{1}{2}\gamma_{jk}^{T}\mathbf{G}_{jk}\gamma_{jk}\right)$ (RIGBY; STASINOPOULOS, 2005).

Para valores fixos de suavização ou hiper-parâmetros λ_{jk}, com $j = 1, 2, \cdots, J_k$ e $k = 1, 2, \cdots, p$ os vetores paramétricos $\boldsymbol{\beta}_k$ e os termos de efeito aleatório $\boldsymbol{\gamma_{jk}}$ são estimados via maximização de uma função verossimilhança penalizada dada por

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_{jk}'\mathbf{G}_{jk}\gamma_{jk}, \tag{2.19}$$

em que $\ell = \sum_{i=1}^{n}\log f\left(y_i \mid \boldsymbol{\theta}^i\right)$ é a função log-verossimilhança.

Para o modelo GAMLSS paramétrico a função log-verossimilhança penalizada (ℓ_p) reduz-se a função log-verossimilhança (ℓ) e os parâmetros β_k, com $k = 1, 2, 3, 4$, são estimados maximizando a função ℓ. Mais detalhes sobre como a função ℓ_p é maximizada podem ser encontrados em Stasinopoulos e Rigby (2007). Utilizando o pacote `gamlss()` em R, é possível avaliar as distribuições da variável dependente e dos diferentes termos aditivos do modelo na estrutura GAMLSS. Note que as classes ML e MLG são dois casos particulares da estrutura GAMLSS (FLORENCIO, 2010).

2.5 Critérios de seleção dos modelos

Na explicação de um fenômeno por meio de qualquer que seja a técnica de regressão há sempre perda de informação. Contudo, o objetivo é obter um modelo no qual esta perda seja mínima, para que o entendimento do problema em estudo não seja comprometido. Um bom modelo é aquele que é capaz de equilibrar a qualidade do ajuste e a complexidade (EMILIANO, 2013).

A seleção de um modelo deve ser baseada em algum critério adequado e bem justificado a respeito de qual é o melhor. Esta deve pautar-se em uma filosofia acerca de modelagem e modelos baseados em inferência estatística, considerando o fato de que as bases de dados são finitas (BURNHAM; ANDERSON, 2004). Além disso, é necessário buscar o modelo mais simples possível e que descreva adequadamente os dados observados nas diversas áreas como agricultura, demografia, ecologia, economia, engenharia, geologia, medicina, entre outros (DEMÉTRIO; CORDEIRO, 2007).

No âmbito da análise de regressão para identificar qual o modelo que mais se aproxima da realidade devem, geralmente, considerar-se princípios científicos com base na teoria da informação. De acordo com Cordeiro e Demétrio (2008) os critérios de informação (IC - *Information Criterion*) de: Akaike (AIC) e Bayesiano (BIC) são duas estatísticas bastante úteis para comparar a qualidade de ajustes dos modelos.

O critério AIC foi introduzido por Akaike (1973). Akaike utilizou a distância de Kullback-Leibler (K-L) visando verificar se um determinado modelo era ou não adequado. É analisada a distância entre o modelo verdadeiro que é algo desconhecido/abstrato e o modelo candidato (AKAIKE, 1974).

O AIC é estimado por

$$AIC = -2\ell + 2k, \tag{2.20}$$

em que ℓ é a função log-verossimilhança que estima a distância K-L e $2k$ é uma estimativa do viés, sendo k o número de parâmetros do modelo estimado.

Embora o AIC seja um estimador assintoticamente não viesado para a distância K-L e que possui também eficiência assintótica, pode não obter bom desempenho para casos em que a amostra é pequena (BUCKLAND; BURNHAM; AUGUSTIN, 1997). Por isso, recomenda-se a utilização deste critério nos casos em que o tamanho da amostra (n) é no mínimo igual a quarenta vezes o número de parâmetros, isto é, $n/k \geq 40$ (ANDERSON; BURNHAM, 2002).

Sugiura (1978) propôs uma correção finita do AIC, ao reanalisar três exemplos de dados estatísticos, para os quais mostrou conclusões razoáveis acerca do AIC para pequenas amostras. Posteriormente, Hurvich e Tsai (1989), derivaram o critério de Akaike Corrigido (AIC_c), que consiste em uma correção de viés para o critério AIC no estudo de regressão e modelos de série temporal autoregressivos para amostras pequenas.

Tanto Akaike (1978) como Schwarz (1978) introduziram o BIC baseando-se em alguns argumentos bayesianos para prová-lo. Schwarz derivou a estatística para modelos de uma família *Koopman-Darmois* e Akaike propôs a estatística para o problema de seleção de variáveis em modelos de regressão linear (AKAIKE, 1978). Assim, o BIC é definido em termos da probabilidade a posteriori, dado por:

$$BIC = -2\ell + k\log(n).$$

O BIC considera que o modelo verdadeiro possui dimensão infinita e, consequentemente, não faz parte do grupo de modelos candidatos. Além disso, este critério também possui consistência e eficiência assintótica, por isso pode não ter um bom desempenho em pequenas amostras, já que estas propriedades assintóticas ficam comprometidas. Como alternativa ao BIC nestes casos, é recomendado a utilização do critério Bayesiano Corrigido (BIC_c), proposto por McQuarrie (1999). Em seu trabalho, McQuarrie (1999), além de derivar a correção do critério de seleção

de Schwarz (BIC), apresenta estudos de simulação para amostras pequenas e em larga escala, incluindo erros não-normais.

Além do AIC e BIC, no caso dos modelos GAMLSS paramétricos, considera-se o critério AIC generalizado (GAIC), obtido pela soma do desvio global ajustado (GD) e uma penalidade fixada (#) para cada grau de liberdade utilizado no modelo. Logo, o GAIC é definido como GAIC(#) = GD + #df, sendo GD o desvio global ajustado dado por GD $= -2\ell(\hat{\boldsymbol{\theta}})$, com $\ell(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \ell(\hat{\boldsymbol{\theta}^i})$ (RIGBY; STASINOPOULOS, 2005). Neste caso, o AIC e o BIC são dois casos particulares do critério GAIC, obtidos quando $\# = 2$ e $\# = \log(n)$, respectivamente.

Para os cinco critérios citados a seleção do melhor modelo dentre uma gama de modelos candidatos, é feita escolhendo-se aquele modelo que obteve o menor valor para o(s) critério(s) considerado(s).

Como ferramenta de diagnóstico considera-se uma coleção de gráficos QQ que constituem o *worm plot* introduzido por Buuren e Fredriks (2001). Este gráfico possibilita visualizar o quanto determinado modelo estatístico se ajusta ao conjunto de dados, permitindo identificar em quais locais o ajuste pode ser melhorado, bem como comparar diferentes modelos ajustados (BUUREN, 2007).

O *worm plot* é uma ferramenta de diagnóstico geral para a análise dos resíduos do modelo, que pode ser usado em conjunto com outros métodos. O eixo vertical contém a diferença entre as medidas de locação teóricas e empíricas (*deviation*). No eixo horizontal são alocados os quantis teóricos da distribuição normal padrão (BUUREN; FREDRIKS, 2001). As curvas elípticas indicam bandas de confiança aproximadas de 95%. Espera-se que os pontos que constituem o *worm* estejam linha horizontal no meio sem forma sistemática e 95% ou mais dos pontos dentro da curva elíptica (STASINOPOULOS; RIGBY; BASTIANI, 2018).

Com base no *worm plot* é possível identificar as regiões (intervalos) de uma covariável, nas quais o modelo não está ajustado adequadamente (STASINOPOULOS; RIGBY; BASTIANI, 2018). Caso nenhuma covariável seja especificada, o *worm plot* é equivalente ao gráfico dos quantis normais dos resíduos sem a tendência (FLORENCIO, 2010). Na Tabela 2.3 é apresentada a maneira de interpretar quatro formatos distintos para a média, variância, assimetria e curtose em um gráfico *worm plot*.

Tabela 2.3 – Interpretação de vários padrões no *worm plot*

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

Fonte: (BUUREN; FREDRIKS, 2001).

Diagnósticos gráficos com base nos resíduos do modelo ajustado são recomendados para verificar a adequabilidade deste ao conjunto de dados (STASINOPOULOS; RIGBY; BASTIANI, 2018). Pereira (2017) destaca que a distribuição dos resíduos quantílicos se aproxima melhor da normalidade do que outros tipos de resíduos. Para a regressão beta por exemplo, o autor verifica que este tipo de resíduo é a melhor alternativa para análise de diagnóstico.

O resíduo quantílico foi proposto por Dunn e Smyth (1996) e é dado por

$$r_i = \phi^{-1}\{F(y_i; \hat{\boldsymbol{\theta}})\},$$

em que ϕ^{-1} corresponde à distribuição normal padronizada e $F(\cdot)$ é um determinada distribuição acumulada a ser usada. O vetor de parâmetros equivalente à posição escala e forma é representado por $\hat{\boldsymbol{\theta}}$.

Rigby (2018) menciona que os resíduos quantílicos são usualmente empregados na avaliação do desempenho das análises de diagnósticos para modelos de regressão mais complexos, como GAMLSS. Além disso, são resíduos que assintoticamente seguem a distribuição normal (STASINOPOULOS; RIGBY; BASTIANI, 2018).

Deste modo, analisa-se o comportamento deste tipo de resíduo. São considerados nos *plots* dos modelos ajustados para cada classe (ML, MLG e GAMLSS), os gráficos dos resíduos *versus* valores ajustados, resíduos *versus* valores observados, densidade dos resíduos e gráfico dos quantis normais.

2.6 A cultura permanente de erva-mate

A utilização da erva-mate como bebida tônica e estimulante já era conhecida e parte integrante dos costumes das primeiras populações que habitaram o sul do Brasil. Os jesuítas, em suas pregações, perceberam que os índios possuíam o hábito de ingerir uma bebida feita das folhas da Ilex Paraguariensis. No início, os jesuítas classificaram a planta como "erva do diabo", proibindo seu uso, pelo efeito estimulante que os deixava irrequietos e muito ativos.

Sem ter sucesso nessa tentativa, passaram a estudar a erva-mate, descobrindo outras maneiras de produção e consumo, fazendo com que suas propriedades espalhassem-se pela Europa, ficando conhecida como “chá dos jesuítas” (SINDIMATE, 2010).

Ao longo dos anos, a erva-mate passou a ser, aos poucos, industrializada e tornar-se um costume das populações que foram atraídas por seu sabor, transformando em um hábito cultural, especialmente no RS, que perdura até os dias atuais (BERGER, 2007; ANTONIAZZI, 2013).

2.6.1 Cadeia produtiva da erva-mate

Denomina-se cadeia produtiva um conjunto de etapas que associam inúmeras fases, que, por sua vez, envolve desde o início do processo de produção até a comercialização do produto final, tais como bens e serviços, chegando, por último na fase de distribuição ao consumidor. Contudo, tratando-se especificamente da cadeia produtiva da erva-mate evidencia-se que ainda não há uma organização que detalhe, exatamente, todos esses processos. Ora por falta de incentivos, orientações, ora pela limitação do conhecimento que possuem os produtores ou inexistência de cooperação entre os mesmos. Por isso, embora o Brasil seja o país responsável pela maior produção da erva-mate, ainda não possui uma estrutura ervateira organizada e sólida que possibilite controlar o estoque do produto para suprir as demandas do mercado (ANTONIAZZI, 2013).

Melo (2010) destaca duas fases da estrutura de produção da erva-mate, sendo a primeira caracterizada pelo processo produtivo da matéria-prima, que envolve desde o fornecimento de insumos externos ao abastecimento e transporte do produto à unidade de industrialização. Enquanto que a segunda fase envolve o processamento industrial da matéria-prima, que compreende a recepção das folhas, o abastecimento da indústria com lenha e demais equipamentos necessários, sendo que destacam-se o procedimento do sapeco da erva-mate, secagem, cancheamento, moagem, empacotamento e, por fim, a comercialização. Tais fases, são especificadas nas Figuras 2.1 e 2.2, a seguir.

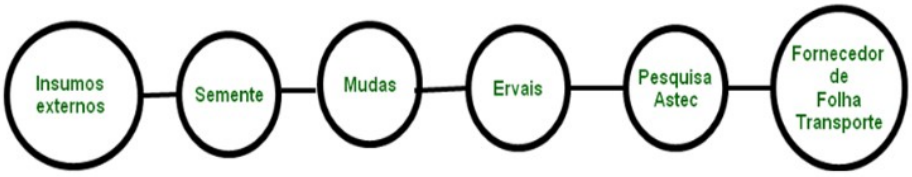

Figura 2.1 – Esquema clássico do processamento da erva-mate na fase I.
Fonte: (MELO, 2010)

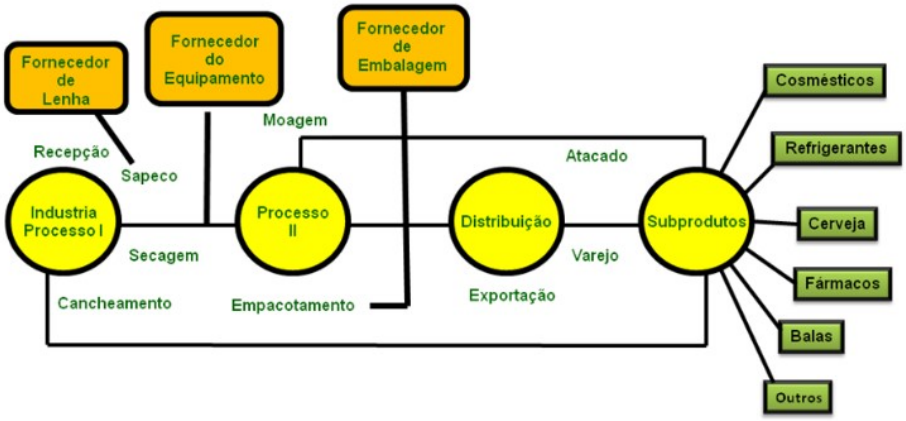

Figura 2.2 – Esquema clássico do processamento da erva-mate na fase II.
Fonte: (MELO, 2010)

Após a chegada da erva-mate em folha verde à agroindústria, em no máximo um dia deve-se iniciar o processo de industrialização de modo com que o produto final não seja prejudicado, quanto a cor e, principalmente, ao sabor (ANTONIAZZI, 2013). Esse processo inicia-se quando a erva-mate em ramos é introduzida em uma grande esteira que a transporta a um cilindro, denominado de sapecadeira, que tal como o nome sugere é responsável pelo processo de sapeco da erva que retira determinado percentual de umidade da folha.

Na sequência, as folhas e galhos são transportados a outra esteira com intuito de esfriar parcialmente a erva-mate, bem como retirar a fumaça existente. Tal esteira conduz a erva sapecada a um cancheador/picador, do qual a erva, parcialmente moída é conduzida a outro cilindro, responsável pela secagem. A erva cancheada sai do secador, com o auxílio de um exaustão que elimina o pó excedente e a conduz ao caracol que realiza a separação das folhas epalitos, por meio de um conjunto de peneiras. Por fim, a erva, já seca e separada, é conduzida à outro caracol que possui bolsas em sua extremidade com capacidade de até 50 kg, as quais armazenam a erva cancheada e o excesso de palitos que são descartados (ANTONIAZZI, 2013).

Por último, após resfriada, a erva cancheada é destinada ao processo de moagem, que se dá por meio de soques ou atritores. De modo geral, se moída em atritores, há novamente a separação de palitos para evitar o comprometimento do sabor suave e a cor verde do produto, conforme já mencionado.

2.6.2 A erva-mate no Brasil e no RS

A erva-mate é uma planta medicinal, cientificamente conhecida como *Ilex Paraguariensis* St. Hil. Este nome científico deve-se ao botânico francês Augusto de Saint-Hilaire que, em 1825, coletou uma amostra da planta, antes conhecida como "erva-do-paraguai" e publicou este nome (BERGER, 2007).

Na América do Sul, a erva-mate é a planta mais utilizada como estimulante, pois possui cafeína na sua composição. As maneiras de uso mais difundidas são o chimarrão (infusão quente) e o tererê (infusão fria). Entretanto, devido às suas propriedades, tem sido frequentemente utilizada na indústria farmacêutica e alimentícia, que tem dado enfoque ao uso de produtos de origem natural (HECK; DE MEJIA, 2007; BERGER, 2007).

A ocorrência desta planta se dá naturalmente na região sul do continente americano e regiões subtropicais. Contudo, aproximadamente, 80% da área de ocorrência pertence ao Brasil, especificamente aos estados do Paraná (PR), Santa Catarina (SC) e RS, ocorrendo também, embora em menor proporção, nos estados de Mato Grosso do Sul (MS), São Paulo (SP), Minas Gerais (MG) e Rio de Janeiro (RJ) (ESMELINDRO et al., 2002; BOGUSZEWSKI, 2007).

São quatro estados brasileiros responsáveis por quase toda a produção nacional: PR, SC, RS e MS. Este último iniciou a produção desta cultura em 1997. O RS aumentou sua área de cultivo, porém reduziu sua participação relativa nas últimas duas décadas. Apesar disso, é o estado que destaca-se na produtividade de erva-mate, conforme pode-se verificar na Figura 2.3 (SINDIMATE, 2016). Estima-se que existam cerca de 250 ervateiras no RS, pertencentes a um mercado com estruturas semelhantes (OLIVEIRA; WAQUIL, 2015).

Em 2014, por exemplo, a erva-mate apresentou o maior valor da produção (R$) de toda a sua série histórica a nível nacional. Com uma produção de 602.484 toneladas em uma área colhida de 70.820 hectares, esta cultura destacou-se com um valor da produção igual a R$ 670.148.000,00. A contribuição do RS a este valor é equivalente a 45,20%, seguido do PR com 42,43% (IBGE, 2016).

O estado do RS produziu a maior quantidade de folha verde de erva-mate a nível nacional, contando com uma produção média de 278.044 toneladas/ano no período de 2013 a 2015, equivalente a 48% da produção nacional (ATLAS, 2009; IBGE, 2016). O estado lidera em termos de consumo e oferta do produto sendo responsável por cerca de 62% da produção brasileira e consumindo 43,6% desta (CHECHI; SCHULTZ, 2016).

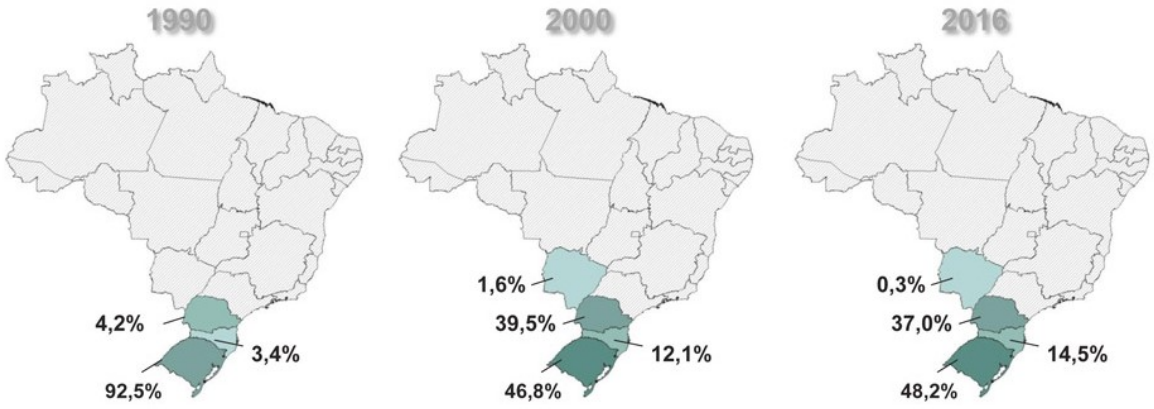

Figura 2.3 – Produção de erva-mate no PR, SC, RS e MS em 1990, 2000 e 2016.
Fonte: SINDIMATE, 2016.

Além disso, o chimarrão, é ritualístico e cultural no RS, sendo a árvore da erva-mate símbolo do estado (BERGER, 2007). Assim, verifica-se o impacto da erva-mate à cultura e economia do estado do RS, dado seu destaque historicamente como atividade produtiva e comercial, e sua extrema relevância cultural (BALZON; SILVA; SANTOS, 2004).

3 MATERIAIS E MÉTODOS

O conjunto de dados utilizado foi obtido do site da Fundação de Economia e Estatística (FEE). A FEE é constituída por um acervo de pesquisas e documentos que reúnem informações de natureza socioeconômica relativos ao estado do RS e cada um de seus municípios, cuja fonte é o Instituto Brasileiro de Geografia e Estatística (IBGE) (FEE, 2016).

Os dados utilizados referem-se à produção de erva-mate do estado do RS no ano de 2016, por ser o último ano que se tinha informação disponível. O estudo concentra-se no valor da produção pago ao produtor rural com a venda de erva-mate em folha verde.

O banco de dados possui cinco variáveis quantitativas contínuas com informações relativas à cultura permanente de erva-mate de 196 municípios gaúchos que obtiveram produção no ano de 2016. O tamanho da amostra considerado é o total de municípios, logo: $n = 196$.

Define-se a variável resposta como:

- VP := valor da produção em R$, em que vp_i é o valor da produção física obtida, considerando-se os preços médios pagos ao produtor no i-ésimo município, para $i = 1, 2, \cdots, 196$.

Denominam-se de covariáveis as demais variáveis do conjunto de dados. São definidas como segue:

- QP := quantidade produzida em toneladas (ton), em que qp_i é a produção obtida no i-ésimo município.
- AC := área colhida em hectares (ha), em que ac_i é parcela da área plantada efetivamente colhida no i-ésimo município.
- ADC := área destinada à colheita em ha, em que adc_i é a área total destinada à colheita equivalente a área ocupada por pés (plantas) em idade produtiva, que tiveram ou não suas produções colhidas no i-ésimo município.
- RM := rendimento médio em Kg/ton, em que rm_i é a razão entre a quantidade produzida e a área colhida no i-ésimo município.

Consequentemente, tanto a variável resposta como as covariáveis são quantitativas contínuas que podem assumir valores pertencentes ao intervalo $(0, \infty)$.

Acrescentaram-se ao banco de dados duas variáveis qualitativas, a saber: polo ervateiro e microrregião referentes aos 196 municípios gaúchos que tiveram produção de erva-mate em

2016. São cinco polos ervateiros existentes no estado do RS (RIGO et al., 2014). De acordo com a divisão geográfica realizada pelo IBGE, o estado gaúcho divide-se em sete mesorregiões, que por sua vez abrangem 35 microrregiões. Destas consideram-se apenas dezoito, referentes aos municípios produtores de erva-mate em 2016.

Acrescentei conforme sugestão RENATA: Os polos ervateiros foram instituídos através de decreto estadual, já que a produção e beneficiamento de erva-mate produto está concentrado em algumas regiões (CHECHI; SCHULTZ, 2017). Estes foram delimitados em 2009 com intuito de racionalizar a gestão da cadeia produtiva da erva-mate no RS, inserida em um cenário de intensidade de trabalho com a atividade e estrutura produtiva estabelecida (MELO, 2016). Ademais, as regiões ervateiras que os constituem foram definidas também por reunirem grande número de viveiros de mudas, indústrias estabelecidas, produtores e volume de produção (EBBING, 2018).

Embora sejam cinco polos ervateiros, define-se uma categoria extra, com intuito de reunir os municípios que não pertencem a algum dos polos. As categorias da variável polo e os municípios que as compõe são os seguintes:

- Alto Taquari (AT): Anta Gorda, Arvorezinha, Coqueiro Baixo, Doutor Ricardo, Fontoura Xavier, Ilópolis, Itapuca, Nova Alvorada, Putinga, Relvado e São José do Herval.
- Alto Uruguai (AU): Aratiba, Áurea, Campinas do Sul, Erebango, Erechim, Gaurama, Getúlio Vargas, Severiano de Almeida e Viadutos.
- Nordeste Gaúcho (NG): Água Santa, Capão Bonito do Sul, Caseiros, Coxilha, Lagoa Vermelha, Machadinho, Maximiliano de Almeida, Paim Filho, Santa Cecília do Sul, São João da Urtiga, São José do Ouro, Tapejara e Tupanci do Sul.
- Planalto Missões (PM): Boa Vista das Missões, Dois Irmãos das Missões, Erval Seco, Novo Barreiro, Palmeira das Missões, São José das Missões, São Pedro das Missões e Seberi.
- Vale do Taquari (VT): Boqueirão do Leão, Cruzeiro do Sul, Gramado Xavier, Herveiras, Mato Leitão, Santa Clara do Sul, Santa Cruz do Sul, Sério, Sinimbu, Vale do Sol e Venâncio Aires.
- Nenhum polo (NP): demais municípios do banco de dados.

As categorias da variável microrregião e os municípios integrantes são:

- Carazinho (Car): Almirante Tamandaré do Sul, Barra Funda, Boa Vista das Missões, Carazinho, Chapada, Coqueiros do Sul, Jaboticaba, Lajeado do Bugre, Nova Boa Vista, Novo Barreiro, Palmeira das Missões, Sagrada Família, Santo Antônio do Planalto, São José das Missões, São Pedro das Missões e Sarandi.

- Caxias do Sul (CS): Bento Gonçalves, Carlos Barbosa, Cotiporã, Fagundes Varela, Santa Tereza e Veranópolis.

- Cruz Alta (CA): Boa Vista do Cadeado, Cruz Alta, Fortaleza dos Valos, Jóia e Saldanha Marinho.

- Erechim (Er): Aratiba, Áurea, Barão de Cotegipe, Barra do Rio Azul, Benjamin Constant do Sul, Campinas do Sul, Carlos Gomes, Centenário, Cruzaltense, Entre Rios do Sul, Erebango, Erechim, Erval Grande, Faxinalzinho, Floriano Peixoto, Gaurama, Getúlio Vargas, Itatiba do Sul, Jacutinga, Marcelino Ramos, Mariano Moro, Paulo Bento, Ponte Preta, Quatro Irmãos, São Valentim, Severiano de Almeida, Três Arroios e Viadutos.

- Frederico Westphalen (FW): Alpestre, Ametista do Sul, Caiçara, Constantina, Cristal do Sul, Dois Irmãos das Missões, Engenho Velho, Erval Seco, Frederico Westphalen, Gramado dos Loureiros, Liberato Salzano, Nonoai, Novo Tiradentes, Novo Xingu, Palmitinho, Rio dos Índios, Rodeio Bonito, Rondinha, Seberi, Taquaruçu do Sul, Três Palmeiras, Trindade do Sul, Vicente Dutra e Vista Alegre.

- Gramado-Canela (GC): Gramado e Santa Maria do Herval.

- Guaporé (Gu): André da Rocha, Anta Gorda, Arvorezinha, Dois Lajeados, Guabiju, Guaporé, Ilópolis, Itapuca, Montauri, Nova Alvorada, Nova Araçá, Nova Bassano, Nova Prata, Paraí, Putinga, São Jorge, São Valentim do Sul, Serafina Corrêa, União da Serra e Vista Alegre do Prata.

- Ijuí (Ij): Ajuricaba, Alegria, Augusto Pestana, Bozano, Chiapetta, Condor, Coronel Barros, Coronel Bicaco, Ijuí, Inhacorá, Panambi, Santo Augusto e São Valério do Sul.

- Lajeado-Estrela (LE): Boqueirão do Leão, Capitão, Coqueiro Baixo, Cruzeiro do Sul, Doutor Ricardo, Encantado, Forquetinha, Marques de Souza, Muçum, Nova Bréscia, Progresso, Relvado, Roca Sales, Santa Clara do Sul, Sério e Vespasiano Correa.

- Nao-Me-Toque: Colorado, Lagoa dos Três Cantos, Não-Me- Toque, Tio Hugo e Victor Graeff.
- Passo Fundo: Água Santa, Casca, Caseiros, Ciríaco, Coxilha, Ernestina, Gentil, Marau, Mato Castelhano, Nicolau Vergueiro, Passo Fundo, Pontão, Ronda Alta, Santa Cecília do Sul, Sertão, Tapejara e Vila Maria.
- Sananduva: Machadinho, Maximiliano de Almeida Paim Filho, São João da Urtiga, São José do Ouro e Tupanci do Sul.
- Santa Cruz do Sul: Gramado Xavier, Herveiras, Mato Leitão, Santa Cruz do Sul, Sinimbu, Vale do Sol e Venâncio Aires.
- Santa Rosa: Alecrim, Novo Machado, Porto Mauá, Santa Rosa, Santo Cristo, Três de Maio, Tucunduva e Tuparendi.
- Santo Ângelo: Bossoroca, Catuípe, Senador, Salgado Filho e Ubiretama.
- Soledade: Barros Cassal, Fontoura Xavier, Ibirapuitã, São José do Herval e Soledade.
- Três Passos: Bom Progresso, Braga, Campo Novo, Doutor Maurício, Cardoso, Horizontina, Humaitá, Nova Candelária, Redentora, São Martinho, Tenente Portela e Três Passos.
- Vacaria: Capão Bonito do Sul, Lagoa Vermelha e Pinhal da Serra.

Para estudar a relação entre o comportamento das covariáveis qualitativas (polo ervateiro e microrregião) e a variável resposta, utilizam-se variáveis *dummy* que representam suas respectivas categorias. Para a covariável polo ervateiro consideram-se, então, cinco variáveis *dummy*, em que a categoria base, no modelo saturado, é a NP. Para a covariável microrregião consideram-se dezessete variáveis *dummy*, tomando como base a categoria Car, também no modelo saturado. Para os modelos finais as categorias base são as *dummy* restantes que não são explícitas na formulação. Assim, o modelo obtido será composto pelo termo constante (intercepto) (GUJARATI, 1970).

Salienta-se que desta forma não há problema de multicolinearidade. Este problema ocorre caso sejam utilizadas todas as variáveis *dummy* no modelo com o intercepto. Somente é possível considerar todas as dummy se no modelo o termo constante é desconsiderado. Logo, esta é outra opção para evitar que haja correlação entre as covariáveis.

As variáveis *dummy* associadas à covariável polo são definidas como segue:

- AT := polo ervateiro Alto Taquari, em que $at_i = 1$ se o i-ésimo município faz parte deste polo e $at_i = 0$, caso contrário.

- AU := polo ervateiro Alto Uruguai, em que $au_i = 1$ se o i-ésimo município faz parte deste polo e $au_i = 0$, caso contrário.

- NG := polo ervateiro Nordeste Gaúcho, em que $ng_i = 1$ se o i-ésimo município faz parte deste polo e $ng_i = 0$, caso contrário.

- PM := polo ervateiro de Planalto Missões, em que $pm_i = 1$ se o i-ésimo município faz parte deste polo e $pm_i = 0$, caso contrário.

- VT := polo ervateiro de Vale do Taquari, em que $vt_i = 1$ se o i-ésimo município faz parte deste polo e $vt_i = 0$, caso contrário.

As variáveis *dummy* para a covariável microrregião são definidas como:

- CS := microrregião de Caxias do Sul, em que $cs_i = 1$ se o i-ésimo município faz parte desta microrregião e $cs_i = 0$, caso contrário.

- CA := microrregião de Cruz Alta, em que $ca_i = 1$ se o i-ésimo município faz parte desta microrregião e $ca_i = 0$, caso contrário.

- Er := microrregião de Erechim, em que $er_i = 1$ se o i-ésimo município faz parte desta microrregião e $er_i = 0$, caso contrário.

- FW := microrregião de Frederico Westphalen, em que $fw_i = 1$ se o i-ésimo município faz parte desta microrregião e $fw_i = 0$, caso contrário.

- GC := microrregião de Gramado-Canela, em que $gc_i = 1$ se o i-ésimo município faz parte desta microrregião e $gc_i = 0$, caso contrário.

- Gu := microrregião de Guaporé, em que $gu_i = 1$ se o i-ésimo município faz parte desta microrregião e $gu_i = 0$, caso contrário.

- Ij := microrregião de Ijuí, em que $ij_i = 1$ se o i-ésimo município faz parte desta microrregião e $ij_i = 0$, caso contrário.

- LE := microrregião de Lajeado-Estrela, em que $le_i = 1$ se o i-ésimo município faz parte desta microrregião e $le_i = 0$, caso contrário.

- NMT := microrregião de Não-Me-Toque, em que $nmt_i = 1$ se o i-ésimo município faz parte desta microrregião e $nmt_i = 0$, caso contrário.

- PF := microrregião de Passo Fundo, em que $pf_i = 1$ se o i-ésimo município faz parte desta microrregião e $pf_i = 0$, caso contrário.

- San := microrregião de Sananduva, em que $san_i = 1$ se o i-ésimo município faz parte desta microrregião e $san_i = 0$, caso contrário.

- SCS := microrregião de Santa Cruz do Sul, em que $scs_i = 1$ se o i-ésimo município faz parte desta microrregião e $scs_i = 0$, caso contrário.

- SR := microrregião de Santa Rosa, em que $sr_i = 1$ se o i-ésimo município faz parte desta microrregião e $sr_i = 0$, caso contrário.

- SA := microrregião de Santo Ângelo, em que $sa_i = 1$ se o i-ésimo município faz parte desta microrregião e $sa_i = 0$, caso contrário.

- So := microrregião de Soledade, em que $so_i = 1$ se o i-ésimo município faz parte desta microrregião e $so_i = 0$, caso contrário.

- TP := microrregião de Três Passos, em que $tp_i = 1$ se o i-ésimo município faz parte desta microrregião e $tp_i = 0$, caso contrário.

- Va := microrregião de Vacaria, em que $va_i = 1$ se o i-ésimo município faz parte desta microrregião e $va_i = 0$, caso contrário.

A estatística descritiva do conjunto de dados foi obtida utilizando o sistema `R` (R Core Team, 2018). Realizou-se uma análise das medidas de posição e dispersão das variáveis quantitativas, bem como o teste de correlação de *Spearman* para estudar o comportamento das covariáveis quantitativas em relação à variável resposta.

Algumas estatísticas foram calculadas para estudar o comportamento das variáveis qualitativas. Estas são sumarizadas em tabelas de frequência complementadas com o acréscimo de relações entre a variável resposta e a produção de erva-mate em cada polo e microrregião. Além disso, gráficos de dispersão e boxplots são apresentados de forma a complementar e facilitar a visualização do comportamento das variáveis que integram o conjunto de dados.

Para a obtenção dos modelos de regressão referentes ao valor da produção de erva-mate utilizou-se a principal função do pacote `gamlss` (original), elaborado por Rigby e Stasinopoulos (2005), denominada também de `gamlss()`, dada a sua facilidade de implementação (STASINOPOULOS; RIGBY, 2007).

Os MLs foram obtidos utilizando a família normal de distribuições na função `gamlss()` e para obter os MLGs foram verificadas distribuições de probalidade candidatas à variável resposta pertencentes à família exponencial. Selecionou-se o modelo probabilístico que apresentou menor valor para o critério de informação de Akaike (AIC) (AKAIKE, 1974) e critério de informação Bayesiano (BIC) (AKAIKE, 1978). Ambos os critérios são descritos no capítulo 2.

Os modelos GAMLSS permitem modelar os outros parâmetros da distribuição condicional da variável dependente, utilizando tanto funções paramétricas como não paramétricas. Isso torna esta classe de modelos bastante flexível (FLORENCIO, 2010). Por isso, para obter os modelos GAMLSS, foram verificadas distribuições de probabilidade para o valor da produção que não pertencem à família exponencial.

Utilizou-se, então, a função `histDist()` do pacote `gamlss` do `R` que possibilita, a partir de uma inspeção visual, identificar a qualidade do ajuste do modelo probabilístico à variável resposta. Esta função, com base no histograma de frequência da variável resposta, ajusta uma distribuição paramétrica candidata entre as que podem adequar-se aos dados. Além disso, fornece os valores para os critérios AIC e BIC. Contudo, não considera a existência das covariáveis (RIGBY; STASINOPOULOS, 2005). Desse modo, verificou-se a adequabilidade das distribuições consideradas conforme os critérios AIC, BIC e critério AIC generalizado (GAIC) dos modelos de regressão ajustados.

Determinam-se apenas modelos aditivos, desconsiderando-se possíveis efeitos de interações entre as covariáveis. Para ambas as classes de modelos ML e MLG foram determinados o modelo completo e, posteriormente, fazendo uso da função `step()` disponível em `R` determinou-se o modelo final para cada classe. A função `step()` seleciona as covariáveis com base nos valores referentes ao critério AIC obtidos mediante um algoritmo que considera o método *stepwise* (R Core Team, 2018).

Na classe dos modelos GAMLSS, para a seleção das covariáveis e obtenção do modelo final foi utilizada a função `stepGAIC.VR()` disponível no pacote `gamlss` (R Core Team, 2018). Esta função é baseada na função `stepAIC()` dada no pacote `MASS`. Mais detalhes podem ser encontrados em Stasinopoulos e Rigby (2007). Destaca-se que os modelos de regressão

determinados em cada classe são comparados pelos referidos critérios de seleção: AIC, BIC e GAIC, geralmente utilizados na literatura.

Por fim, como ferramenta de análise de diagnóstico de resíduos dos modelos determinados são considerados o *worm plot*, gráfico dos resíduos quantílicos *versus* valores ajustados e *versus* valores observados, gráfico da densidade estimada dos resíduos e gráfico normal Q-Q *plot*. Ressalta-se que considerou-se o nível de significância de 5% para julgar significativas ou não as covariáveis nos modelos obtidos.

4 RESULTADOS E DISCUSSÃO

Neste capítulo é apresentada a análise descritiva do banco de dados considerado, tais como: mínimo, máximo, mediana, média, quartis, desvio padrão, coeficiente de variação, assimetria e curtose. São apresentados gráficos para a variável resposta e cada covariável quantitativa para observar os comportamentos destas. Posteriormente são apresentados e discutidos os modelos saturados e finais obtidos de acordo com cada técnica abordada no capítulo 2.

4.1 Estatística descritiva do conjunto de dados

Tabela 4.1 – Medidas de posição e dispersão

Estatística	**Variável**				
	VP	**QP**	**AC**	**ADC**	**RM**
Mínimo	4.000,00	7,00	1,00	1,00	180,00
Primeiro quartil	26.250,00	30,00	4,00	5,00	7.500,00
Mediana	77.500,00	107,50	12,00	15,00	10.000,00
Média	1.071.974,49	1.516,03	156,22	170,33	9.414,26
Terceiro quartil	258.750,00	401,25	50,00	50,00	10.000,00
Máximo	52.345.000,00	66.000,00	6.600,00	7.300,00	20.000,00
Desvio padrão	5.332.347,25	6.883,14	696,35	760,53	3.090,26
Coeficiente de variação	497,43	454,03	445,74	446,51	32,83
Assimetria	8,56	8,01	8,12	8,12	0,78
Curtose	79,61	71,80	73,61	73,64	4,25

De acordo com as medidas descritivas apresentadas na Tabela 4.1, percebe-se que a variável VP abrange um expressivo intervalo de valores, sendo o mínimo R$ 4.000,00 e o máximo R$ 52.345.000,00. O terceiro quartil indica que 75% dos municípios gaúchos produtores no ano de 2016 obtiveram um valor da produção inferior a R$258.750,00. O histograma da variável resposta (VP) dado na Figura 4.1 complementa e corrobora o exposto na Tabela 4.1.

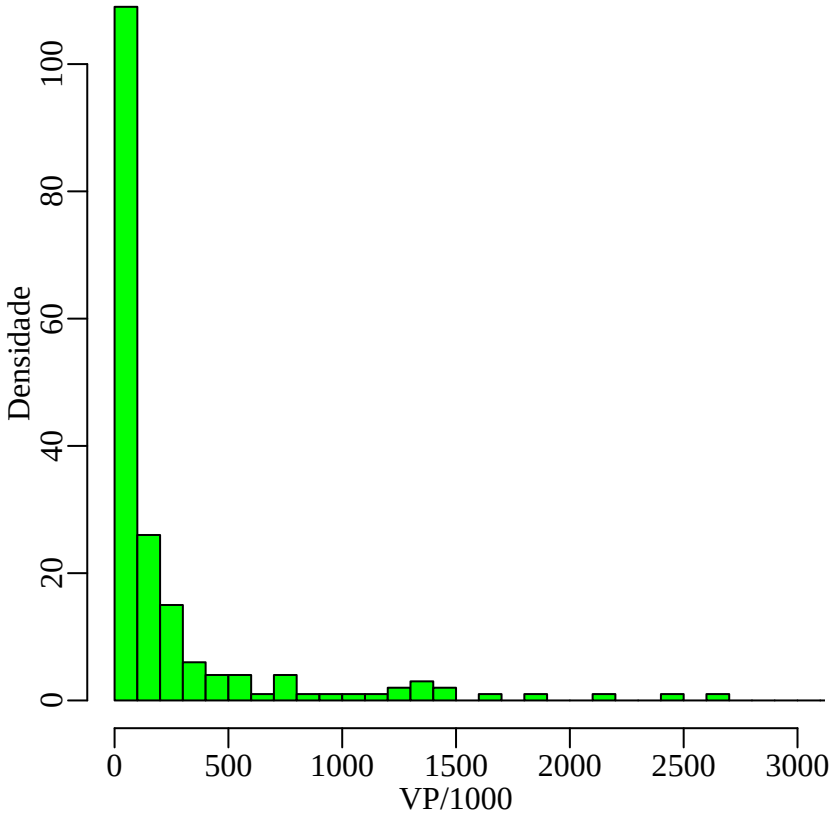

Figura 4.1 – Histograma da variável resposta

A maior parte dos valores assumidos por esta variável são similares, porém há uma quantidade de *outliers* equivalentes a 5,61% das observações que consistem em elevados valores da produção, superestimando a média da variável conforme verificado na Tabela 4.1.

A covariável RM apresenta a menor variabilidade, quanto a análise do coeficiente de variação (32,83%), bem como possui média e mediana muito próximas, destacando-se consequentemente, com a menor assimetria igual a 0,78. As demais variáveis (VP, QP, ADC e AC) possuem dispersão extremamente elevada, com coeficientes de variação acima de 445%. Assim, a mediana, embora robusta, é preferível à média para caracterizar a tendência central destas variáveis.

Além disso, evidencia-se que o coeficiente de assimetria é positivo para todas as variáveis, indicando caudas maiores à direita. Analogamente, todas possuem curtose maior que zero, logo a distribuição da variável resposta e das covariáveis é leptocúrtica, isto é, possui caudas pesadas.

Os gráficos da Figura 4.2 mostram que todas as variáveis apresentam *outliers*. As covariáveis ADC e AC, porém, apresentam quantidades de *outliers* superiores as demais. Tal

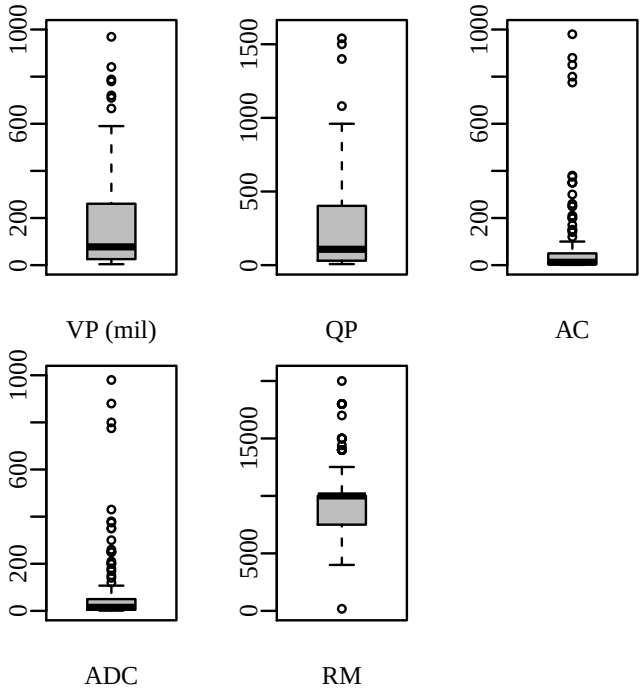

Figura 4.2 – Boxplots da variável resposta e covariáveis quantitativas

como evidenciado na Tabela 4.1, a covariável que possui a menor dispersão é o RM. Com exceção desta, todas apresentam comportamento assimétrico como indicado pelo coeficiente de assimetria.

A Figura 4.3 mostra os gráficos de dispersão das variáveis duas a duas para os dados observados. É possível perceber que existe uma relação mais fraca entre a variável resposta e a covariável RM em comparação com as demais covariáveis quantitativas. Com exceção do RM, percebe-se uma relação linear positiva entre o VP e a QP, AC e ADC.

A Tabela 4.2 mostra o resultado do teste de correlação de *Spearman* realizado para verificar a significância destas relações lineares. As covariáveis QP, AC e ADC apresentam correlação positiva e significativa com a variável VP, sendo os coeficientes de correlação iguais a 0,98, 0,94 e 0,93, respectivamente. Já o RM apresenta correlação fraca com o valor de produção, conforme também pode ser verificado na Figura 4.3, porém significativa.

Tabela 4.2 – Teste de correlação de *Spearman*

	$\rho(\cdot,\cdot)$	**p-valor**
(VP, QP)	0,98	< 0,0001
(VP, AC)	0,94	< 0,0001
(VP, ADC)	0,93	< 0,0001
(VP, RM)	0,24	0,0006

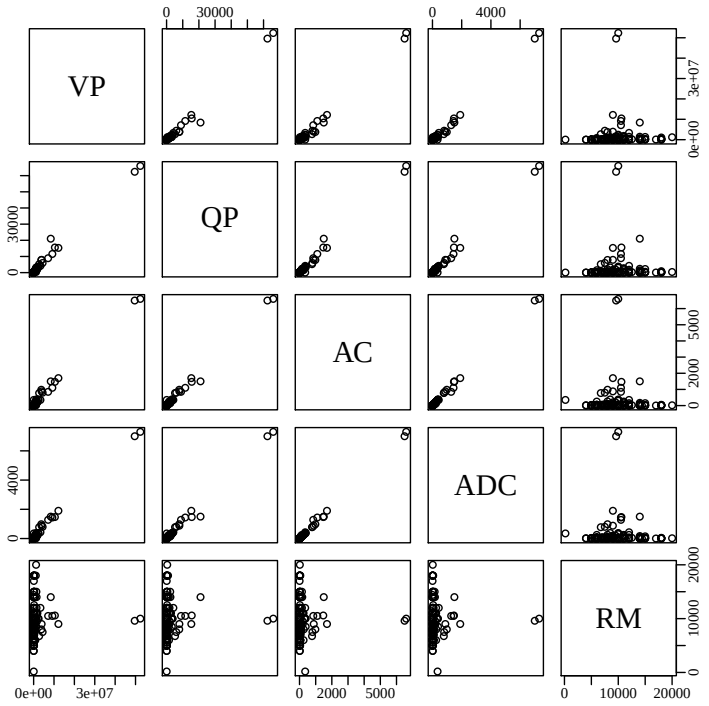

Figura 4.3 – Gráfico de dispersão

Acrescentei conforme sugestão RENATA. A correlação positiva e significativa entre a variável resposta e as covariáveis QP, AC e ADC deve-se ao fato da natureza destas variáveis. Conforme definidas no capítulo 3 verifica-se que, de fato, há uma relação linear entre estas e a variável resposta. É evidente que o preço médio pago ao produtor rural aumentará à medida que este produzir mais. Para isso, provavelmente destinará maior área para a colheita, e consequentemente terá maior área colhida.

Na Tabela 4.3 são apresentadas algumas estatísticas que expressam o comportamento da variável resposta de acordo com os polos ervateiros. Também apresenta-se o percentual da contribuição relativa ao total da produção em cada polo. Verifica-se que a maior parte dos municípios não pertence a algum polo ervateiro, totalizando 144 municípios com 12,60% da produção.

O polo AT se destaca com 62,53% da produção total e com o maior valor de produção obtido igual a R$ 145.509.000,00, contando com um total de 11 municípios que o integram. De acordo com Ebbing (2018) este polo responde por aproximadamente 60% da produção estadual, pois a ele pertencem os dois municípios gaúchos de maior produção: Ilópolis e Arvorezinha. O destaque do polo AT é confirmado na Figura 4.4.

Tabela 4.3 – Estatísticas referentes aos polos ervateiros

Polo	$\mathbf{f}_{obs}$	$\mathbf{f}_{rel}$**(%)**	**VP (mil)**	**QP (%)**
AT	11,00	5,61	145.509,00	62,53
AU	9,00	4,59	14.960,00	9,40
NG	13,00	6,64	4.514,00	1,99
PM	8,00	4,08	16.331,00	10,74
VT	11,00	5,61	4.239,00	2,75
NP	144,00	73,47	24.554,00	12,60
Total	**196,00**	**100,00**	**210.107,00**	**100,00**

Por outro lado, embora o polo VT tenha produzido mais que o polo NG, obteve o menor valor de produção na faixa de R$ 4.239.000,00. Os polos AT e AU não apresentam *outliers*, enquanto os demais quatro polos possuem o valor da produção influenciado pela presença de observações discrepantes, principalmente, o grupo que reúne os municípios que não pertencem a algum polo ervateiro.

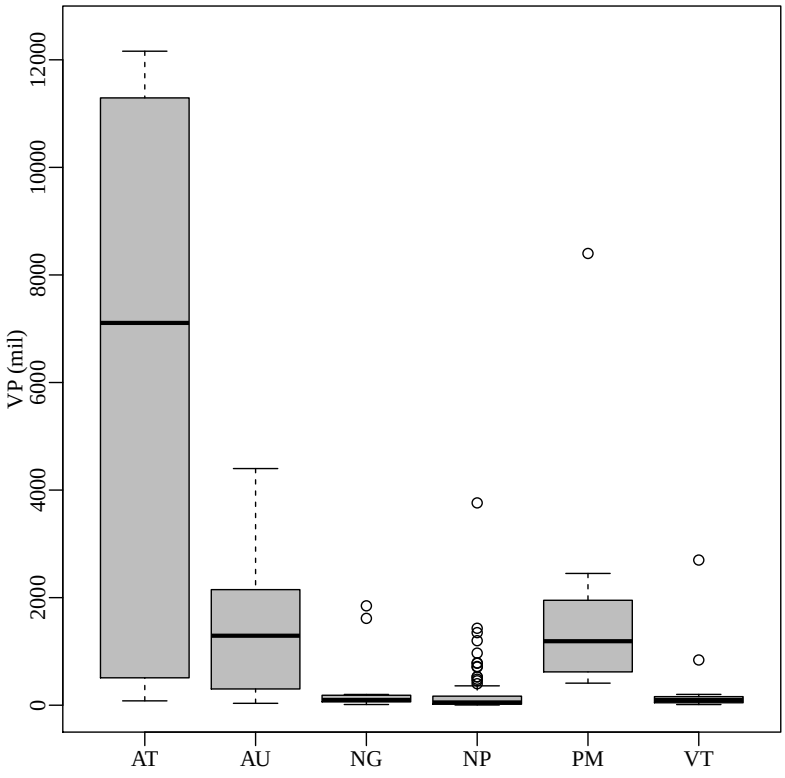

Figura 4.4 – Boxplot do VP *versus* polos ervateiros

De forma análoga à análise descritiva do comportamento das variáveis qualitativas referentes aos polos ervateiros, a Tabela 4.4 apresenta o comportamento da variável resposta de acordo com as covariáveis qualitativas que se referem as microrregiões às quais pertencem os

municípios produtores.

Tabela 4.4 – Estatísticas referentes às microrregiões

Microrregião	$\mathbf{f}_{obs}$	$\mathbf{f}_{rel}$**(%)**	**VP (mil)**	**QP (%)**
Car	16,00	8,16	15.035,00	10,20
CS	6,00	3,06	264,00	0,14
CA	5,00	2,55	227,00	0,06
Er	28,00	14,29	22.674,00	15,02
FW	24,00	12,24	4.340,00	2,06
GC	2,00	1,02	1.296,00	0,15
Gu	20,00	10,20	136.492,00	57,87
Ij	13,00	6,63	2.992,00	1,09
LE	16,00	8,16	2.039,00	0,96
NMT	5,00	2,55	92,00	0,04
PF	17,00	8,67	2.522,00	1,16
San	6,00	3,06	2.407,00	1,10
SCS	7,00	3,57	4.065,00	2,64
SR	8,00	4,08	319,00	0,19
SA	4,00	2,04	1.103,00	0,49
So	5,00	2,55	11.339,00	5,68
TP	11,00	5,61	1.083,00	0,39
Va	3,00	1,53	1.818,00	0,74
Total	**196,00**	**100,00**	**210.107,00**	**100,00**

Com o maior valor de produção, a microrregião de Guaporé que reúne 10,20% municípios produtores, possui também o maior percentual de produção igual a 57,87%. Por outro lado, o menor valor da produção e a menor quantidade produzida refere-se a microrregião de Não-Me-Toque, da qual apenas cinco municípios fazem parte. A microrregião de Erechim é a que reúne o maior número de municípios sendo este igual a 28, equivalente a 14,29% do total. Esta produz a segunda maior quantidade equivalente a 15,02%, bem como obteve o segundo maior valor de produção igual a R$ 22.674.000,00.

Na Figura 4.5 é expressivo o alto valor da produção obtido na microrregião de Guaporé, bem como a alta dispersão apresentada por esta covariável em relação à variável resposta. Com exceção das covariáveis CA, CS, GC, NMT, SA, So, TP e Va, todas as demais possuem *outliers*.

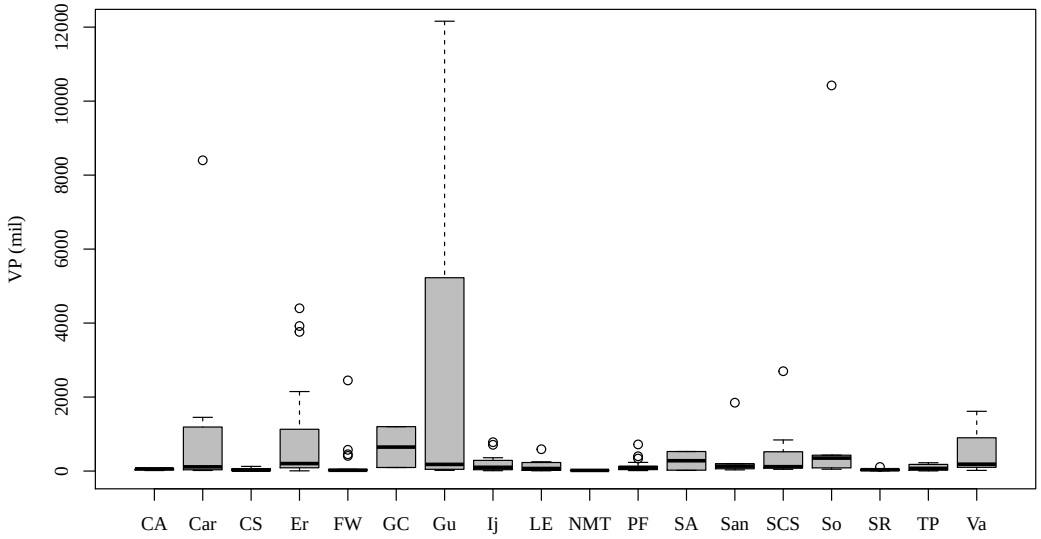

Figura 4.5 – Boxplot do VP *versus* microrregiões

4.2 Modelo linear para o valor da produção de erva-mate

Ajustou-se o modelo linear clássico para o valor da produção como função de todas as covariáveis descritas na seção 3, obtendo-se o modelo saturado, denominado de modelo 11. Mediante o uso da função `step()` determinou-se o modelo final (modelo 12) desta técnica de regressão. Ambos os modelos 11 e 12 são mostrados na Tabela 4.5.

Na Tabela 4.6 são apresentados os valores obtidos para os critérios de seleção AIC e BIC referentes aos modelos 11 e 12. Comparando-os percebe-se que os menores valores para ambos os critérios referem-se ao modelo 12.

Tabela 4.6 – Valores do AIC e BIC para os modelos 11 e 12

Critérios	**Modelo 11**	**Modelo 12**
AIC	5.692,46	5.663,28
BIC	5.784,25	5.699,34

Na Figura 4.6 verifica-se que a curva que acompanha os valores centrais do gráfico *worm plot* extrapola os limites elípticos, o que indica elevados desvios e um pobre ajuste do modelo aos dados. Além disso, evidencia-se pelo eixo vertical do *worm plot* que os pontos extrapolam também o limite $(-3,\ 3)$.

Tabela 4.5 – ML saturado e final para o valor da produção da erva-mate no RS em 2016.

Modelo	Covariável	Estimativa	Erro Padrão	p-valor
11	Intercepto	-235060,51	37364,15	< 0,0001
	QP	126,33	13,20	< 0,0001
	AC	-1.357,42	283,18	< 0,0001
	ADC	7.151,47	237,19	< 0,0001
	RM	22,21	3,04	< 0,0001
	AT	-205.721,37	47.498,22	< 0,0001
	AU	-446.595,03	41.036,52	< 0,0001
	NG	37.747,75	44.417,91	0,3966
	PM	-292.667,33	44.894,55	< 0,0001
	VT	70.811,50	58.837,26	0,2305
	CS	10.307,46	49.220,89	0,8344
	CA	25.897,66	52.795,03	0,6244
	Er	-271.264,47	36.063,65	< 0,0001
	FW	59.378,20	32.950,97	0,0733
	GC	468.678,50	77.644,53	< 0,0001
	Gu	-128.456,01	38.554,80	0,0011
	Ij	74.979,43	39.395,55	0,0587
	LE	-77.442,38	41.323,21	0,0626
	NMT	39.093,77	52.454,96	0,4571
	PF	23.509,12	37.796,50	0,5348
	San	22.385,84	66.469,30	0,7367
	SCS	-525.772,06	76.037,68	< 0,0001
	SR	118,86	45025,84	0,9979
	SA	58.776,29	57.164,22	0,3053
	So	-283.410,18	57.432,36	< 0,0001
	TP	-37.345,25	44.488,72	0,4024
	Va	89.903,69	69.005,52	0,1944
12	Intercepto	-212.322,53	115.237,70	0,0670
	QP	116,38	48,90	0,0183
	ADC	5.943,39	441,46	< 0,0001
	RM	20,31	11,38	0,0760
	AU	-457.951,29	175.057,76	0,0096
	PM	-271.082,03	175.148,88	0,1234
	Er	-287.301,60	107.617,82	0,0083
	GC	472.920,08	311.195,82	0,1303
	SCS	-478.606,85	169.813,44	0,0053
	So	-400.067,87	198.051,03	0,0448

Similarmente, pode ser observada a falta de ajuste do modelo 12 na Figura 4.7. No QQ-*plot* os resíduos afastam-se fortemente da linha dos quantis teóricos da distribuição normal. No gráfico da densidade estimada verifica-se o mesmo evidenciado no gráfico anterior e que não são normalmente distribuídos. Os demais gráficos mostram que os resíduos não parecem ter comportamento aleatório.

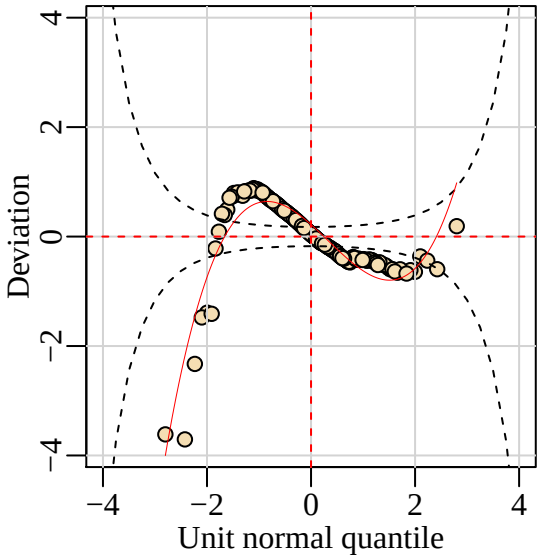

Figura 4.6 – *Worm plot* do modelo 12

A homocedasticidade e normalidade dos resíduos foi verificada pelos testes de Breusch-Pagan e Jarque-Bera, respectivamente. Ao nível de 5% rejeitaram-se as hipóteses nula de que estes possuem variância constante e são normalmente distribuídos. Desse modo, esta forma funcional para a predição do valor da produção de erva-mate não é adequada.

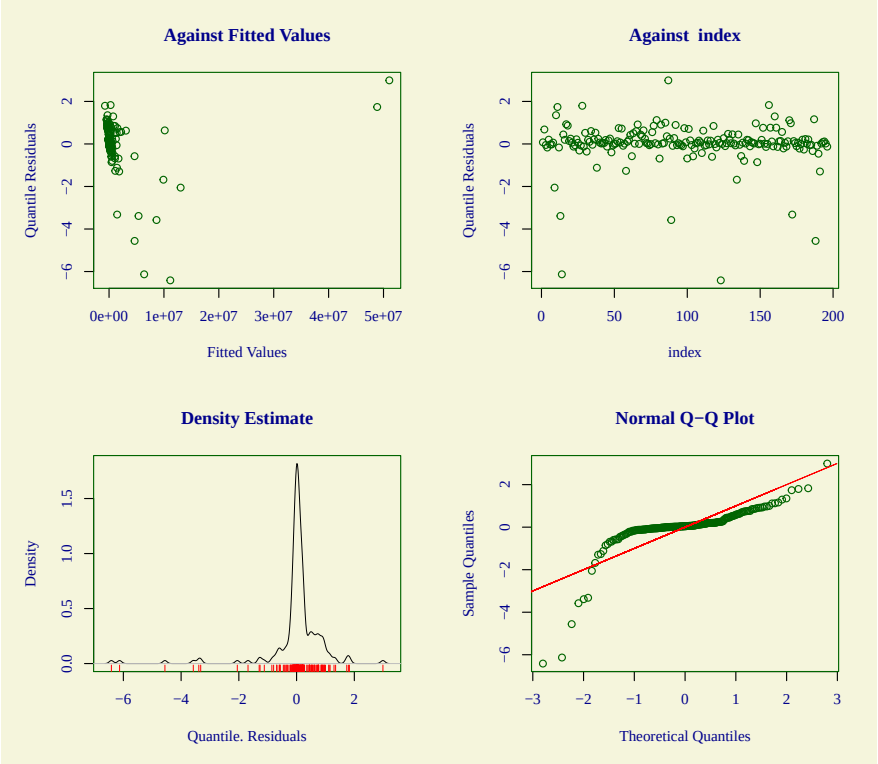

Figura 4.7 – Gráficos de resíduos do modelo 12

A baixa qualidade do ajuste do modelo 12 deve-se, principalmente, ao fato de a variável VP não seguir uma distribuição normal. Contudo, de acordo com Dantas e Cordeiro (2000) é natural haver ausência de normalidade de uma variável resposta relacionada à preços. Isso justifica-se por este tipo de variável ter domínio nos reais positivos, ao passo que a distribuição normal possui suporte no conjunto dos números reais.

4.3 Modelo linear generalizado para o valor da produção de erva-mate

Para o ajuste do MLG avaliou-se a possibilidade da variável resposta ter distribuição Exponencial, Gama ou Normal Inversa devido ao fato de assumir apenas valores reais positivos. A melhor descrição do valor da produção foi obtida no MLG com a distribuição Exponencial e função de ligação canônica ($\log$). O ajuste do MLG final (modelo 22) para a predição do valor da produção de erva-mate no RS é sumarizado na Tabela 4.7, juntamente com o modelo saturado (21), e é dado por

$$\begin{aligned} g(VP) = {} & \beta_0 + \beta_1 QP + \beta_2 AC + \beta_3 RM + \beta_4 AT + \beta_5 AU + \beta_6 PM + \beta_7 VT \\ & + \beta_8 CS + \beta_9 CA + \beta_{10} Er + \beta_{11} FW + \beta_{12} NMT + \beta_{13} San + \beta_{14} SCS \\ & + \beta_{15} SR + \beta_{16} TP + \beta_{17} Va, \qquad \text{(Modelo 22)} \end{aligned}$$

em que $VP \sim Exponencial(\theta)$ e $\eta = \log(\mu)$.

Diferentemente do modelo 12, todas as covariáveis são significativas ao nível de 5% de significância no modelo 22. Neste percebe-se que as covariáveis que apresentaram efeito negativo ao valor médio da produção de erva-mate são: QP, VT, CS, CA, FW, NMT, SR e TP. O sinal negativo da covariável QP indica que há uma redução no valor de VP ao passo que a quantidade produzida de erva-mate aumenta. Já o sinal negativo da covariável polo indica que se o munícipio pertence ao polo VT tende a obter menor valor da produção de erva-mate. Analogamente interpreta-se o sinal negativo para as covariáveis referentes às microrregiões: CS, CA, FW, NMT, SR, TP.

O *worm plot* para o modelo 22 é dado na Figura 4.8. Embora o comportamento deste gráfico tenha sido melhor que o anterior referente ao modelo 12, neste caso o *worm* também ultrapassa as bandas de 95% de confiança. O *worm* se apresenta inclinado para cima à esquerda no gráfico, o que indica que a curtose está levemente acentuada.

Tabela 4.7 – MLG saturado e final para o valor da produção da erva-mate no RS em 2016.

Modelo	Covariável	Estimativa	Erro Padrão	p-valor
21	Intercepto	9,7251	0,3734	< 0,0001
	QP	-0,0002	0,0001	0,0538
	AC	0,0028	0,0006	< 0,0001
	ADC	0,0001	0,0015	0,9237
	RM	0,0002	0,0000	< 0,0001
	AT	1,8038	0,4289	< 0,0001
	AU	1,1243	0,4070	0,0064
	NG	-0,4738	0,4442	0,2876
	PM	2,2415	0,4476	< 0,0001
	VT	-0,8286	0,5882	0,1607
	CS	-1,0107	0,4921	0,0415
	CA	-1,0886	0,5277	0,0407
	Er	0,9592	0,3586	0,0082
	FW	-0,7353	0,3289	0,0267
	GC	0,1315	0,7762	0,8657
	Gu	0,4638	0,3823	0,2268
	Ij	0,4006	0,3936	0,3103
	LE	-0,0930	0,4124	0,8219
	NMT	-1,7206	0,5244	0,0013
	PF	0,3180	0,3777	0,4010
	San	1,7412	0,6644	0,0096
	SCS	2,2060	0,7560	0,0040
	SR	-1,2324	0,4500	0,0068
	SA	0,4399	0,5715	0,4425
	So	0,6430	0,5468	0,2412
	TP	-1,4278	0,4442	0,0016
	Va	1,9664	0,6897	0,0049
22	Intercepto	9,8806	0,3521	< 0,0001
	QP	-0,0002	0,0001	0,0222
	AC	0,0034	0,0011	0,0020
	RM	0,0002	0,0000	< 0,0001
	AT	1,9567	0,4055	< 0,0001
	AU	1,0938	0,4110	0,0085
	PM	2,0753	0,4694	< 0,0001
	VT	-1,1673	0,5129	0,0240
	CS	-1,2607	0,4243	0,0034
	CA	-1,3453	0,4627	0,0041
	Er	0,6787	0,2765	0,0151
	FW	-0,9559	0,2499	0,0002
	NMT	-1,9629	0,4627	< 0,0001
	San	1,0136	0,4629	0,0298
	SCS	2,2445	0,6506	0,0007
	SR	-1,4901	0,3731	0,0001
	TP	-1,7326	0,3799	< 0,0001
	Va	1,2368	0,5895	0,0373

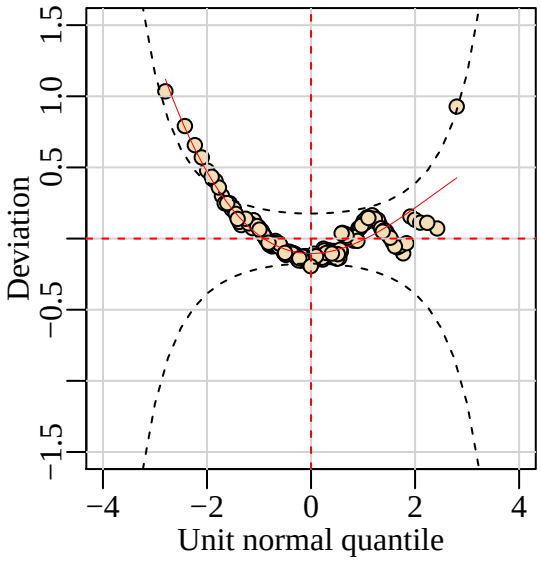

Figura 4.8 – *Worm plot* do modelo 22

Nos gráficos da Figura 4.9 de diagnóstico dos resíduos do modelo 22 é possível notar que o ajuste é razoável quando comparado ao modelo 12. Os valores obtidos para os critérios de seleção considerados são apresentados na Tabela 4.8

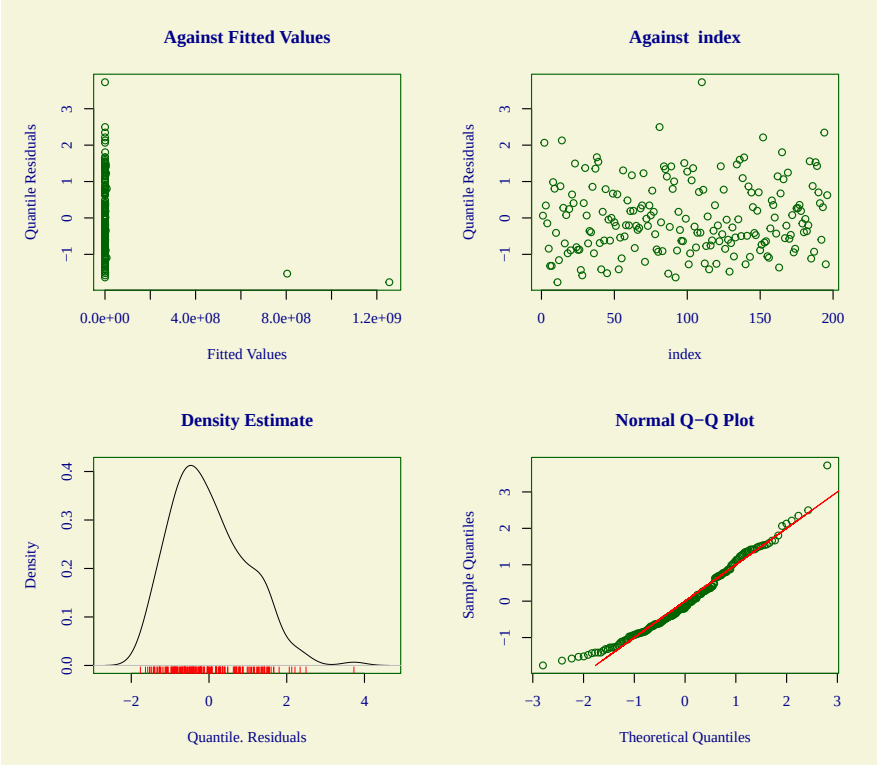

Figura 4.9 – Gráficos de resíduos do modelo 22

Tabela 4.8 – Valores do AIC e BIC para os modelos 21 e 22

Critérios	Modelo 21	Modelo 22
AIC	5.181,25	5.167,17
BIC	5.269,76	5.226,17

Houve redução dos valores de AIC e BIC para ambos os modelos 21 e 22 em relação aos modelos 11 e 12, obtidos via técnica de regressão linear clássica. Isso indica que a qualidade deste ajuste é superior à anterior. Assim, o uso da distribuição exponencial para a variável resposta contribuiu, levemente, na melhoria do ajuste do modelo ao conjunto de dados. Entretanto, é necessário maior flexibilidade para modelagem do valor da produção de erva-mate.

4.4 Modelo aditivo generalizado para locação escala e forma para o valor da produção de erva-mate

Entre as distribuições apresentadas na Tabela 2.2, cujo suporte são os reais positivos, são plausíveis os modelos: Box-Cox, Cole e Green (BCCT), Box-Cox-t (BCT), *Skew* t tipo 3 (ST3), *Johnson's* original SU (JSUo) e *Shash* (SHASH) para modelar a variável resposta referente ao valor da produção de erva-mate. Porém, destacaram-se dentre estes, os modelos BCCT e BCT. Evidencia-se este fato com base na Figura 4.10, que apresenta o histograma da variável resposta ajustada a cada uma destas distribuições.

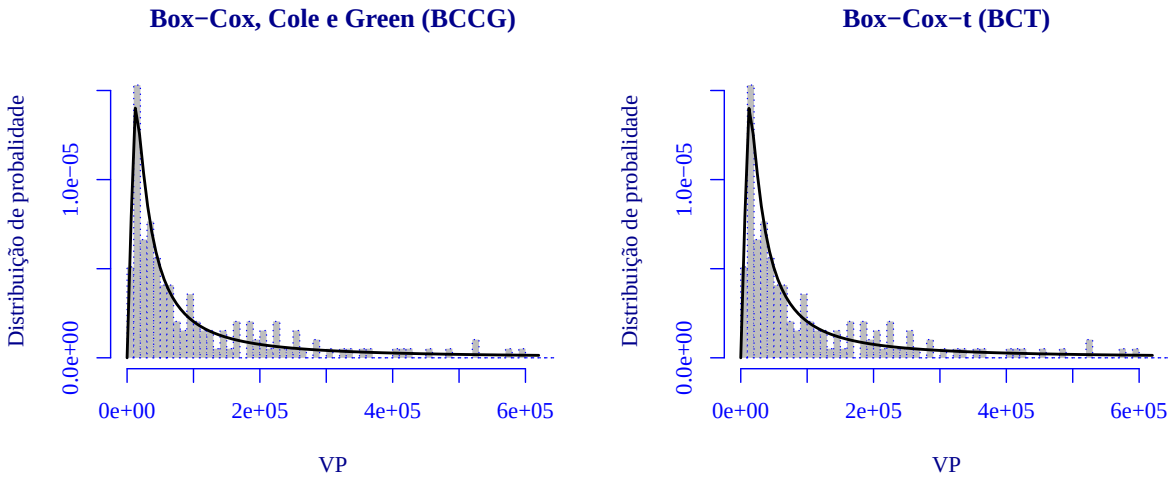

Figura 4.10 – Ajustes das distribuições Box-Cox, Cole e Green e Box-Cox-t à variável resposta.

Na Tabela 4.9 são dados os valores obtidos para cada critério de acordo com os modelos

probabilísticos candidatos.

Tabela 4.9 – Valores do AIC e BIC para as distribuições BCCG e BCT

Distribuição	**AIC**	**BIC**
BCCG	5.298,87	5.308,71
BCT	5.292,85	5.313,97

Contudo, apenas a inspeção visual dos histogramas apresentados na Figura 4.10 e a análise dos valores obtidos para os critérios AIC e BIC não é suficiente para decidir qual modelo melhor se ajusta à variável resposta, já que não são consideradas as covariáveis existentes no conjunto de dados. Faz-se necessário então, analisar o desempenho dos modelos de regressão obtidos com base nas distribuições de probablidade que melhor se ajustam à variável dependente (FLORENCIO, 2010).

Por isso apresenta-se no Apêndice A uma tabela comparativa entre os valores dos critérios AIC, BIC e GAIC para os modelos GAMLSS, considerando-se as distribuições candidatas mencionadas para a variável resposta. A qualidade do ajuste do modelo em que o valor da produção tem distribuição BCCG e BCT é superior à dos demais modelos GAMLSS (tanto saturado como final) com as outras distribuições.

Embora a distribuição BCT forneça um ajuste razoável para a variável resposta, o modelo BCCG destaca-se apresentando os menores valorees para os critérios de seleção considerados. O *worm plot* dos resíduos, assim como os gráficos de diagnósticos também apresentam melhor comportamento. Consequentemente, discutem-se apenas os modelos ajustados considerando que a variável VP segue o modelo probabilístico BCCG.

A distribuição BCCG é amplamente utilizada em estimação por curvas de percentis de referência no método LMS (COLE; GREEN, 1992). É um modelo adequado para modelar dados com assimetria positiva ou negativa. Se $Y \sim$ BCCG (μ, σ, ν) então os parâmetros são modelados via GAMLSS da seguinte forma

$$\mu = s_1(x)$$
$$\log \sigma = s_2(x)$$
$$\nu = s_3(x),$$

em que $0 < y < \infty$, $0 < \mu < \infty$ é o parâmetro de locação equivalente a mediana, $0 < \sigma < \infty$ é o parâmetro de escala aproximadamente dado pelo coeficiente de variação e $\infty < \nu < \infty$ representa o parâmetro de assimetria (STASINOPOULOS; RIGBY; BASTIANI, 2018). Nota-

se que como a mediana é por si própria resistente à presença de *outliers*, estes terão menor influência na modelagem da variável resposta (ROCKE, 1989). Mais detalhes sobre o modelo probabilístico BCCG podem ser encontrados em Rigby *et. al* (2017).

Considerando então que a variável resposta segue uma distribuição BCCG (μ, σ, ν), ajustam-se os modelos saturado e final, modelos 31 e 32, respectivamente, para os três parâmetros equivalentes à locação, escala e forma da variável VP. Apresenta-se o modelo 31 na Tabela 4.10. Salienta-se que consideram-se as funções de ligação padrão para os parâmetros de locação, escala e forma, conforme a Tabela 2.2.

Tabela 4.10 – Modelo 31 para o valor da produção de erva-mate

Parâmetro	**Covariável**	**Estimativa**	**Erro Padrão**	**p-valor**
μ	Intercepto	7062,03	2914,34	0,0164
	QP	614,80	20,20	< 0,0001
	AC	-1.064,66	446,20	0,0181
	ADC	1.084,38	440,35	0,0148
	RM	0,06	0,23	0,7941
	AT	8.638,47	19.928,78	0,6652
	AU	-9.609,59	8.421,70	0,2555
	NG	-1.200,51	3.438,73	0,7274
	PM	135.028,83	69.099,14	0,0523
	VT	-419,87	2.574,93	0,8707
	CS	-7.821,82	2.663,87	0,0038
	CA	4.821,59	3.850,71	0,2123
	Er	-11.542,57	2.453,97	< 0,0001
	FW	-5.658,19	2.107,35	0,0080
	GC	72.621,01	24.229,34	0,0031
	Gu	-7.357,33	3.924,33	0,0626
	Ij	2.234,97	2.977,62	0,4539
	LE	-6.514,85	2.498,23	0,0099
	NMT	-2.536,53	2.763,67	0,3600
	PF	-2.081,67	3.057,74	0,4969
	San	-19.982,48	12.587,44	0,1143
	SCS	-28.473,25	8.589,94	0,0011
	SR	-7.935,94	2.312,40	0,0007
	SA	339,38	3.294,20	0,9181
	So	-2.167,74	16.225,88	0,8939
	TP	-7.223,30	2.162,54	0,0010
	Va	579,81	4.208,06	0,8906
σ	Intercepto	-1,39	0,05	< 0,0001
ν	Intercepto	-0,84	0,25	0,0009

Dada a flexibilidade permitida pela classe dos modelos GAMLSS quanto ao ajuste de distribuições de probabilidade para a variável resposta, outros modelos foram avaliados para o

valor da produção. Porém, a qualidade do ajuste dos modelos de regressão obtidos foi inferior ao ajuste considerando a distribuição BCCG para o variável VP, conforme valores dos critérios de seleção considerados (AIC e BIC). No apêndice A apresenta-se uma tabela comparativa entre os ajustes conformes critérios AIC e BIC.

Na modelagem do parâmetro de locação, os coeficientes das covariáveis que mostraram-se estatisticamente significativos ao nível de 5% são: Car e NP (categorias base equivalentes ao intercepto, QP, AC, ADC, CS, Er, FW, GC, LE, SCS, SR e TP.

Mediante o uso da função `stepGAIC.VR`, as covariáveis são selecionadas de acordo com o critério GAIC obtendo-se o modelo 32 apresentado na Tabela 4.11.

Tabela 4.11 – Modelo 32 para o valor da produção de erva-mate

Parâmetro	**Covariável**	**Estimativa**	**Erro Padrão**	**p-valor**
μ	Intercepto	9.170,97	1.228,20	< 0,0001
	QP	610,94	19,18	< 0,0001
	AC	-1.166,46	405,30	0,0045
	ADC	1.174,28	404,99	0,0042
	PM	137.988,91	68.613,31	0,0458
	CS	-9.240,58	2.176,52	< 0,0001
	Er	-13.372,39	1.725,68	< 0,0001
	FW	-7.079,02	1.391,88	< 0,0001
	GC	70.670,30	23.793,55	0,0034
	Gu	-8.977,43	3.519,28	0,0115
	LE	-8.264,27	1.779,31	< 0,0001
	NMT	-3.973,24	2.341,65	0,0915
	PF	-4.263,21	2.186,52	0,0528
	San	-22.290,85	12.108,12	0,0673
	SCS	-29.658,83	8.160,69	0,0004
	SR	-9.318,97	1.707,93	< 0,0001
	TP	-8.785,97	1.435,22	< 0,0001
σ	Intercepto	-1,38	0,06	< 0,0001
ν	Intercepto	-0,92	0,23	< 0,0001

O modelo 32 também pode ser dado por

$$\begin{aligned}
\mu = \; & \beta_0 + \beta_1\, QP + \beta_2\, AC + \beta_3\, ADC + \beta_4\, PM + \beta_5\, CS + \beta_6\, Er + \beta_7\, FW + \beta_8\, GC + \\
& \beta_9\, Gu + \beta_{10}\, LE + \beta_{11}\, NMT + \beta_{12}\, PF + \beta_{13}\, San + \beta_{14}\, SCS \\
& + \beta_{15}\, SR + \beta_{16}\, TP \\
\log(\sigma) = \; & \beta_0 \\
\nu = \; & \beta_0,
\end{aligned}$$

em que a variável valor da produção (VP) segue uma distribuição BCCG com parâmetro de posição (μ), de escala (σ) e de assimetria (ν).

No modelo 32, na modelagem do parâmetro de locação são significativas a 5% as covariáveis NP e Car (intercepto), QP, AC, ADC, CS, Er, FW, GC, Gu, LE, SCS, SR, TP. Examinam-se os valores de cada estimativa dos parâmetros referentes a estas covariáveis que foram significativas como segue.

- Pela análise do termo constante verifica-se que o valor da produção de erva-mate mediano é de R$ 9.170,97 que é inferior à mediana geral da variável resposta R$ 77.500,00 dada na Tabela 4.1.

- A estimativa do parâmetro associado à covariável QP indica que a cada unidade de mudança em QP o valor da produção tem um aumento de R$ 9.781,91, quando as demais covariáveis são mantidas fixas.

- A cada unidade de alteração na variável AC, a resposta mediana do valor da produção aumenta R$ 8.004,51, considerando-se as demais variáveis constantes. Corrobora-se isso analisando os resultados do estudo de correlação apresentado na Tabela 4.2 e Figura 4.3.

- O valor mediano da produção aumenta em R$ 10.345,25 à cada unidade de alteração na variável ADC. Isto também explicado devido a correlação positiva e significativa entre esta e a variável resposta.

- Dentre os polos ervateiros o único significativo é polo PM. Este possui efeito positivo em relação à mediana da variável resposta. Corrobora-se esta contribuição via análise da Tabela 4.3. O polo PM contribui com 10,74% da quantidade produzida de erva-mate e possui o terceiro maior valor da produção.

- Na microrregião Caxias do Sul (CS) o valor mediano da produção de erva-mate reduz-se em R$ 69,61. A Figura 4.5 corrobora esta informação, na qual é possível verificar o baixo valor da produção nesta microrregião.

- Nos municípios pertencentes às microrregiões de Erechim, Frederico Westphalen, Guaporé, Lajeado-Estrela, Não-Me-Toque, Passo Fundo, Santiago, Santa Cruz do Sul, Santa Rosa e Três Passos o valor da produção diminui.

- O sinal positivo da covariável GC indica que o valor da produção aumenta à medida em que a localização dos municípios pertence à microrregião de Gramado-Canela. Esta é a única covariável dentre as microrregiões com efeito positivo à variável resposta. Isto pode ser explicado por ser a que contém o menor número de municípios produtores integrantes, conforme Tabela 4.5.

A análise das estimativas dos parâmetros referentes às covariáveis indica que as variações da variável VP relacionam-se com o polo PM e as microrregiões do estado gaúcho. Isso se deve ao fato de que há particularidades nos sistemas de produção e nas estruturas de governança estabelecidas com intuito de comercializar o produto (PICOLOTTO et al., 2013). Ademais, o polo PM possui a segunda maior produção estadual, contando com cerca de 15% da produção gaúcha, com destaque para o município de Palmeira das Missões, ao qual pertence aproximadamente 7,1% da produção estadual e 4,9% da área colhida (**?**).

O *worm plot* do modelo 32, Figura 4.11, apresentou o melhor comportamento, pois não há pontos fora das bandas de confiança como nos demais *worms*. Os pontos centrais do *worm* estão muito próximos da origem, indicando que a mediana está bem ajustada. Verifica-se também que a variância, assimetria e curtose foram bem modeladas.

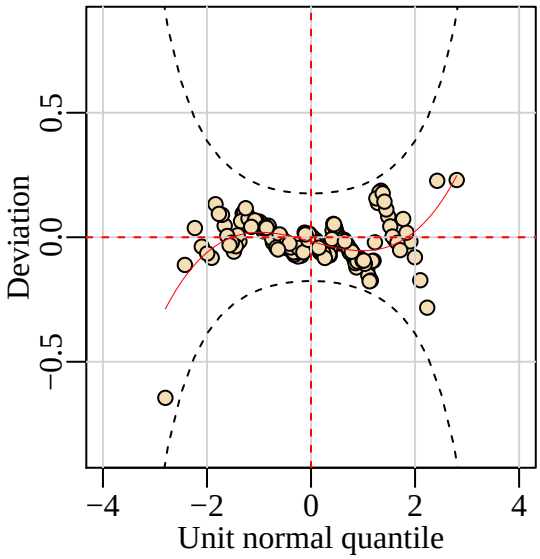

Figura 4.11 – *Worm plot* do modelo 32

Os gráficos de diagnóstico dos resíduos são dados na Figura 4.12. No gráfico dos resíduos *versus* valores ajustados não é possível notar que o comportamento dos resíduos é aleatório em torno de zero com variância constante, devido à escala do gráfico. Por isso, apresenta-se o

mesmo gráfico na Figura 4.13, o qual indica bom ajuste do modelo e variância homocedástica.

O gráfico dos resíduos *versus* observações indica comportamento aleatório destes. Ademais, verifica-se a normalidade razoável dos resíduos quantílicos conforme pode-se notar nos dois últimos gráficos. Logo, este diagnóstico indica um bom ajuste do modelo 32.

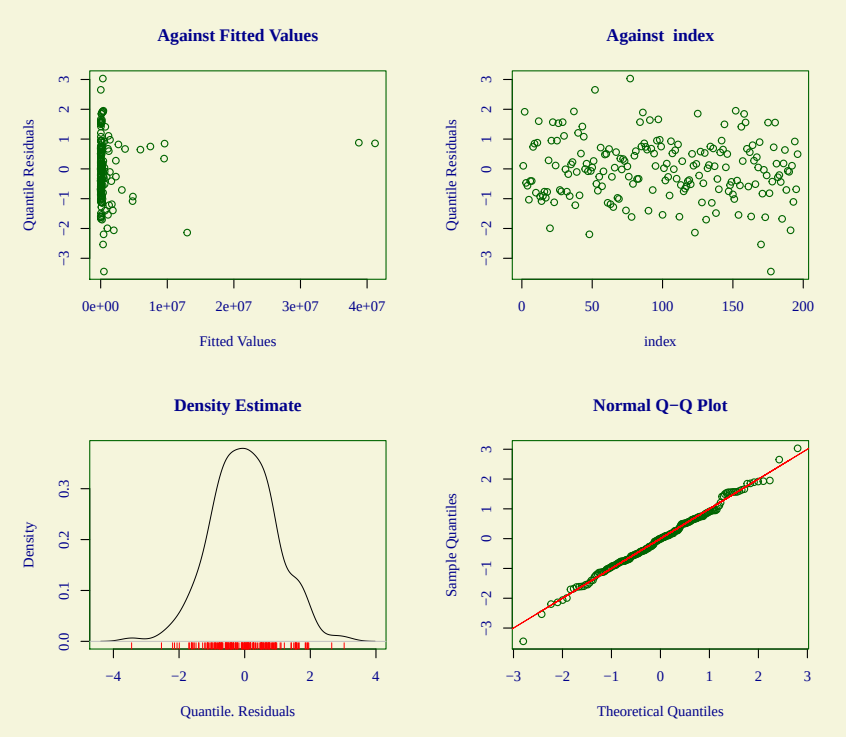

Figura 4.12 – Gráficos de resíduos do modelo 32

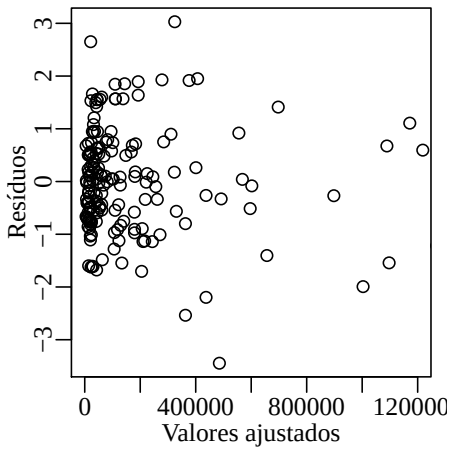

Figura 4.13 – Gráfico dos resíduos *versus* valores ajustados do modelo 32

Os valores obtidos para os três critérios de seleção considerados na escolha do modelo

GAMLSS, são dados na Tabela 4.12. Evidencia-se os menores valores para os três no modelo 32.

Tabela 4.12 – Valores do AIC e BIC para os modelos 31 e 32

Critérios	**Modelo 31**	**Modelo 32**
AIC	4.604,53	4.588,16
BIC	4.699,60	4.650,44
GAIC	4.657,89	4.623,12

Salienta-se que a interpretação das estivamativas dos parâmetros do modelo 32 pode ser estendida para os demais modelos ajustados (11,12, 21 e 22), considerando-se as restrições das variáveis que não mostraram-se significativas. Porém optou-se por apresentar estas conclusões somente ao modelo 32, já que obteve a melhor qualidade de ajuste.

4.5 Comparação dos modelos finais ajustados

Ao comparar os *worm plots* dos modelos finais apresentados nas Figuras 4.6, 4.8 e 4.11 evidencia-se a melhora na qualidade do ajuste à medida que aumenta a complexidade da técnica de regressão considerada. O mesmo é possível observar nas Figuras 4.7, 4.9 e 4.12 referentes ao diagnóstico dos resíduos. De forma a corroborar com tais análises gráficas de diagnóstico, comparam-se os valores obtidos para os critérios de seleção considerados na Tabela 4.13.

Tabela 4.13 – Comparação entre os modelos finais estimados via classe dos ML, MLG e GAMLSS

Modelo	**Classe**	**AIC**	**BIC**
12	ML	5.663,28	5.699,34
22	MLG	5.167,17	5.226,17
32	GAMLSS	4.588,16	4.650,44

O critério AIC no modelo 32 é aproximadamente 18,99% menor que no modelo 12. Analogamente o critério BIC é 18,40% menor no modelo 32 se comparado ao modelo 12. Isto indica a melhora da qualidade do ajuste do modelo mediante esta última técnica, uma vez que possibilitou maior flexibilidade à modelagem do valor da produção de erva-mate no estado gaúcho. Portanto, o modelo 32 parece ser o mais indicado para predizer o valor da produção de erva-mate no RS em 2016.

5 CONCLUSÃO

A principal finalidade da realização desta monografia foi obter um modelo de regressão que explicasse a variação do valor da produção da erva-mate no estado do RS em 2016. Foram obtidos seis modelos, distribuídos dois a dois nas três classes de modelos estatísticos consideradas: ML, MLG e GAMLSS. Para cada técnica de regressão foram obtidos os modelos saturado e final de acordo com o método *stepwise* de seleção de covariáveis.

Obtiveram-se estatísticas descritivas do banco de dados e identificaram-se as covariáveis que explicam a variabilidade do valor de produção nos diferentes municípios gaúchos. Assim, os parâmetros de todos os modelos ajustados foram estimados, sendo discutidas apenas as estimativas referentes ao modelo final que obteve a melhor qualidade de ajuste.

A análise estatística descritiva evidencia que a variável resposta possui elevada assimetria positiva e presença de *outliers*. Verifica-se correlação positiva entre as variáveis quantitativas geradoras do valor da produção. A maioria das covariáveis qualitativas são significativas na explicação do valor da produção de erva-mate para todos os modelos.

Com o pressuposto de normalidade da variável resposta não satisfeito, verifica-se a não adequabilidade do ajuste dos MLs ao conjunto de dados. Pela análise de diagnóstico dos resíduos nota-se que as demais suposições como normalidade e independência dos resíduos também não foram atendidas. No *worm plot* da Figura 4.6 é evidente a baixa qualidade do ajuste.

O mesmo pode ser verificado na Tabela 4.6, que apresenta os maiores valores para os critérios AIC e BIC. Dado que a distribuição normal não é capaz de modelar assimetria e/ou curtose, justifica-se o modelo com baixa qualidade de ajuste, o qual pode levar a conclusões distorcidas da realidade.

Da família exponencial de distribuições o modelo probabilístico que melhor se ajusta à variável resposta é a distribuição exponencial. Embora a qualidade do ajuste do modelo 22 referente à classe dos MLGs seja superior ao modelo 11 obtido, a suposição de normalidade dos resíduos também é violada. O *worm plot* da Figura 4.8 referente ao modelo 22 indica curtose acentuada e verifica-se que os pontos centrais ultrapassam os limites de confiança, logo o modelo ajustado não é adequado.

Quanto ao modelo final GAMLSS (modelo 32) evidencia-se sua superioridade de ajuste em relação as demais classes. Tanto os valores relativos aos critérios de seleção dos modelos (AIC, BIC e GAIC) quanto a análise gráfica dos resíduos e *worm plot* do modelo 32 confirmam

a boa qualidade do ajuste. Destaca-se ainda a possibilidade de modelar os três parâmetros da distribuição BCCG, já que a classe GAMLSS permite descrever também os outros parâmetros da distribuição da variável resposta em termos das covariáveis.

Ademais, identificam-se quais covariáveis quantitativas e qualitativas inluenciam no aumento ou redução do valor da produção da erva-mate no RS conforme interpretação das estimativas do modelo 32. O polo PM, por exemplo, integra todos os modelos obtidos sendo significativo e possuindo efeito positivo na maioria dos ajustes. Justifica-se este fato por ser o polo que conta com a segunda maior produção de erva-mate no RS, depois do polo AT. Estas conclusões, são, portanto, muito valiosas no que tange entender os fatores que interferem na variabilidade do valor da produção desta cultura permanente tão relevante para o desenvolvimento econômico, principalmente do estado gaúcho.

REFERÊNCIAS

AKAIKE, H. A. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v.19, n.6, p.716–723, 1974.

AKAIKE, H. A. Bayesian analysis of the minimum AIC procedure. **Annals of the Institute of Statistical Mathematics A**, v.30, n.1, p.9–14, 1978.

ANDERSON, D. R.; BURNHAM, K. P. Avoiding pitfalls when using information-theoretic methods. **The Journal of Wildlife Management**, p.912–918, 2002.

ANDREWS, D. F. A note on the selection of data transformations. **Biometrika**, v.58, n.2, p.249–254, 1971.

ANTONIAZZI, M. S. **A cadeia produtiva da erva-mate no município de Três Passos:** produção, industrialização e comercialização. 2013.

ATLAS SOCIOECONÔMICO DO RIO GRANDE DO SUL. Porto Alegre: SCP, 2009.

BALZON, D. R.; SILVA, J. C. G. L. da; SANTOS, A. J. dos. Aspectos Mercadológicos de Produtos Florestais Não Madeireiros Análise Retrospectiva. **Floresta**, v.34, n.3, 2004.

BARTLETT, M. S. The Use of Transformations. **Biometrics**, v.3, n.1, p.39–52, 1947.

BERGER, G. D. C. Modelos volumétricos para erva-mate (*Ilex paraguariensis* A. St.-Hil.), na região nordeste do estado do Rio Grande do Sul: uma análise através das técnicas de regressão. **Ciência e Tecnologia de Alimentos**, 2007.

BERKSON, J. Application of the logistic function to bio-assay. **Journal of the American Statistical Association**, v.39, n.227, p.357–365, 1944.

BICKEL, P. J.; DOKSUM, K. A. An analysis of transformations revisited. **Journal of the american statistical association**, v.76, n.374, p.296–311, 1981.

BIRCH, M. Maximum likelihood in three-way contingency tables. **Journal of the Royal Statistical Society. Series B (Methodological)**, p.220–233, 1963.

BLISS, C. I. The calculation of the dosage-mortality curve. **Annals of Applied Biology**, v.22, n.1, p.134–167, 1935.

BOGUSZEWSKI, J. H. **Uma história cultural da erva-mate**: o alimento e suas representações. 2007. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Paraná.

BOX, G. E. P.; COX, D. R. An Analysis of Transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, v.26, n.2, p.211–252, 1964.

BUCKLAND, S. T.; BURNHAM, K. P.; AUGUSTIN, N. H. Model Selection: an integral part of inference. **Biometrics**, v.53, n.2, p.603–618, 1997.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding aic and bic in model selection. **Sociological methods & research**, v.33, n.2, p.261–304, 2004.

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

BUUREN, S. van. Worm plot to diagnose fit in quantile regression. **Statistical Modelling**, v.7, n.4, p.363–376, 2007.

CARROLL, R. J. A Robust Method for Testing Transformations to Achieve Approximate Normality. **Journal of the Royal Statistical Society. Series B (Methodological)**, v.42, n.1, p.71–78, 1980.

CHARNET, R. et al. **Análise de modelos de regressão linear**: com aplicações. 2.ed. Campinas, SP: Unicamp, 2008.

CHECHI, L. A.; SCHULTZ, G. A produção de erva mate: um estudo da dinâmica produtivas nos estados do sul do brasil. **Enciclopédia Biosfera, Goiânia**, v.13, n.23, p.16–26, 2016.

CHECHI, L. A.; SCHULTZ, G. Arranjos produtivos locais de erva-mate no sul do Brasil: caracterização das organizações processadoras e relações estabelecidas. **Espacios**, v.38, n.32, p.14, 2017.

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

CORDEIRO, G. M.; DEMÉTRIO, C. G. **Modelos lineares generalizados e extensões**. São Paulo, v.33, 2008.

DANTAS, R.; CORDEIRO, G. Uma avaliação do mercado de apartamentos do Recife utilizando modelos lineares generalizados. In: 19º **Congresso Panamericano de Avaliações.** 2000.

DEMÉTRIO, C.; CORDEIRO, G. Modelos lineares generalizados. **Simpósio de Estatística Aplicada à Experimentação Agronômica** , v.12, 2007.

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

DYKE, G. V.; PATTERSON, H. D. Analysis of Factorial Arrangements when the Data are Proportions. **Biometrics**, v.8, n.1, p.1–12, 1952.

EBBING, M. Diagnóstico da Cadeia Produtiva da Erva-Mate no estado do Rio Grande do Sul. **IBRAMATE - Instituto Brasileiro da Erva-Mate**, n.01, 2018.

EMILIANO, P. C. Critérios de informação: como eles se comportam em diferentes modelos. **Doutoramento em Estatıstica e Experimentaçao Agropecuária**, 2013.

ESMELINDRO, M. C. et al. Caracterização físico-química da erva-mate: influência das etapas do processamento industrial. **Ciência e Tecnologia de Alimentos**, v.22, n.2, p.193–204, 2002.

FEIGL, P.; ZELEN, M. Estimation of exponential survival probabilities with concomitant information. **Biometrics**, p.826–838, 1965.

FLORENCIO, L. **Engenharia de avaliações com base em modelos GAMLSS**. 2010. 125 p. 2010. Dissertação (Mestrado em Estatıstica). Departamento de Estatıstica, Universidade Federal de Pernambuco.

FUNDAÇÃO DE ECONOMIA E ESTATÍSTICA. **Culturas Permanentes**. 2016.

GALTON, F. I. Family likeness in stature. **Proceedings of the Royal Society of London**, v.40, n.242-245, p.42–73, 1886.

GLASSER, M. Exponential survival with covariance. **Journal of the American Statistical Association**, v.62, n.318, p.561–568, 1967.

GUJARATI, D. Use of dummy variables in testing for equality between sets of coefficients in linear regressions: a generalization. **The American Statistician**, v.24, n.5, p.18–22, 1970.

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica**. 5.ed. Porto Alegre: AMGH, 2011. 924p.

HASTIE, T. J.; TIBSHIRANI, R. J. **Generalized additive models**.Chapman & Hall, London, 1990.

HECK, C. I.; DE MEJIA, E. G. Yerba Mate Tea (*Ilex paraguariensis*): a comprehensive review on chemistry, health implications, and technological considerations. **Journal of food science**, v.72, n.9, p.R138–R151, 2007.

HERNANDEZ, F.; JOHNSON, R. A. The Large-Sample Behavior of Transformations to Normality. **Journal of the American Statistical Association**, v.75, n.372, p.855–861, 1980.

HURVICH, C. M.; TSAI, C.-L. Regression and time series model selection in small samples. **Biometrika**, v.76, n.2, p.297–307, 1989.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agrícola Municipal**. Brasil, 1996.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agrícola Municipal**. Brasil, 2016.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agrícola Municipal**. Brasil, 2017.

LIN, X.; ZHANG, D. Inference in generalized additive mixed modelsby using smoothing splines. **Journal of the royal statistical society: Series b (statistical methodology)**, v.61, n.2, p.381–400, 1999.

LINDSEY, J. K. **Applying generalized linear models**. New York: Springer, 1997.

MAIA, A. G. **Econometria**: conceitos e aplicações. [S.l.]: Saint Paul Editora, 2019.

MCCULLAGH, P.; NELDER, J. **Generalized linear models**. 2.ed. London: Chapman and Hall, 1989.

MCCULLOCH, C. E. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. **Journal of the American Statistical Association**, v.92, n.437, p.162–170, 1997.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, Linear, and Mixed Models**. 1.ed. United States of America: John Wiley & Sons, 2001. 358p.

MCQUARRIE, A. D. A small-sample correction for the Schwarz SIC model selection criterion. **Statistics & probability letters**, v.44, n.1, p.79–86, 1999.

MELO, I. Mapeamento da cadeia produtiva da erva-mate no municipio de Machadinho: desafios e propostas. 2010. 48 p. **Monografia (Especialização em Gestão do Agronegocio)– Universidade do Vale dos Sinos, Novo Hamburgo**, 2010.

MELO, I. B. d. Os Polos ervateiros do RS: distribuição geográfica. **EMATER (Regional Passo Fundo)**, 2016.

NELDER, J. A. Inverse Polynomials, a Useful Group of Multi-Factor Response Functions. **Biometrics**, v.22, n.1, p.128–141, 1966.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General)**, v.135, n.3, p.370–384, 1972.

OLIVEIRA, S. V. d.; WAQUIL, P. D. Dynamics of production and commercialization of yerba mate in Rio Grande do Sul, Brazil. **Ciência Rural**, v.45, n.4, p.750–756, 2015.

PEARSON, K.; LEE, A. On the laws of inheritance in man: i. inheritance of physical characters. **Biometrika**, v.2, n.4, p.357–462, 1903.

PEREIRA, G. H. On quantile residuals in beta regression. **Communications in Statistics-Simulation and Computation**, p.1–15, 2017.

PICOLOTTO, P. et al. A dinâmica de produção e de comercialização da erva-mate nos cinco polos ervateiros do estado do Rio Grande do Sul. **I Seminário de Jovens Pesquisadores em Economia e Desenvolvimento**, 2013.

R Core Team. **R**: a language and environment for statistical computing. Brisbane, Austrália: R Foundation for Statistical Computing, 2018.

RA FISHER, M. On the mathematical foundations of theoretical statistics. **Phil. Trans. R. Soc. Lond. A**, v.222, n.594-604, p.309–368, 1922.

RASCH, G. **Studies in mathematical psychology:** Probabilistic models for some intelligence and attainment tests, 1960.

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

RIGBY, R. et al. Distributions for modelling location, scale, and shape: using gamlss in R. **Communications in Statistics-Simulation and Computation**, p.1–15, 2017.

RIGBY, R.; STASINOPOULOS, D. The GAMLSS project: a flexible approach to statistical modelling. In: **New Trends in Statistical Modelling: proceedings of the 16th International Workshop on Statistical Modelling**. 2001. v.337, p.345.

RIGO, L. et al. Análise do mercado da erva-mate no Brasil e no Rio Grande do Sul. **Área Temática - D. Estudos setoriais, cadeias produtivas, sistemas locais de produção**, 2014.

ROCKE, D. M. Robust control charts. **Technometrics**, v.31, n.2, p.173–184, 1989.

SCHNEIDER, P. R.; SCHNEIDER, P. S. P.; SOUZA, C. A. M. d. **Análise de Regressão Aplicada à Engenharia Florestal**. 2.ed. Santa Maria: Facos, 2009.

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

SINDICATO DA INDÚSTRIA DO MATE NO ESTADO DO RIO GRANDE DO SUL. Dados Estatísticos. Porto Alegre, 2010.

SINDICATO DA INDÚSTRIA DO MATE NO ESTADO DO RIO GRANDE DO SUL. **Dados Estatísticos**. Porto Alegre, 2016.

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.; RIGBY, R. A.; BASTIANI, F. D. GAMLSS: a distributional regression approach. **Statistical Modelling**, v.18, n.3-4, p.248–273, 2018.

STORCK, L. et al. Precisão experimental em erva-mate (*Ilex paraguaiensis* St. Hil.). **Ciência Florestal**, v.12, n.1, p.159–161, 2002.

SUGIURA, N. Further analysts of the data by Akaike' s information criterion and the finite corrections. **Communications in Statistics - Theory and Methods**, v.7, n.1, p.13–26, 1978.

TURKMAN, A. A.; SILVA, G. L. **Modelos Lineares Generalizados - da teoria à prática**. Lisboa: [s.n.], 2000.

VOGT, G. A.; NEPPEL, G.; SOUZA, A. M. de. A atividade ervateira no Planalto Norte Catarinense: a indicação geográfica como alternativa para a (re) valorização do produto erva-mate. **DRd-Desenvolvimento Regional em debate**, v.6, n.2, p.64–87, 2016.

ZANIN, V.; MEYER, L. G. Evolução da margem de comercialização da erva mate no Rio Grande do Sul. **Revista IPecege**, v.4, n.1, p.7–18, 2018.

ZIPPIN, C.; ARMITAGE, P. Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. **Biometrics**, p.665–672, 1966.

APÊNDICES

APÊNDICE A – Tabela comparativa entre os modelos GAMLSS estimados com outras distribuições para a variável resposta

Tabela A.1 – Valores do AIC e BIC para outros modelos GAMLSS

Distribuição	**Modelo**	**AIC**	**BIC**	**GAIC**
BCCG	Saturado	4604.53	4699.60	4657.89
	Final	4.588,16	4.650,44	4623.12
BCT	Saturado	4.607,70	4.706,10	4.662,90
	Final	4.592,41	4.654,70	4.627,37
LOGNO	Saturado	5160,09	5251,88	5211,61
	Final	5145,15	5204,15	5178,27
GG	Saturado	5161,13	5256,25	5214,49
	Final	5145,76	5208,08	5180,72
BCTo	Saturado	5163,23	5261,63	5218,43
	Final	5147,88	5213,48	5184,68