UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

RONDINELLY DUARTE DE OLIVEIRA JÚNIOR

SELEÇÃO DE VARIÁVEIS EXPLICATIVAS NOS MODELOS ADITIVOS GENERALIZADOS DE LOCAÇÃO, ESCALA E FORMA

RECIFE
2024

RONDINELLY DUARTE DE OLIVEIRA JÚNIOR

SELEÇÃO DE VARIÁVEIS EXPLICATIVAS NOS MODELOS ADITIVOS GENERALIZADOS DE LOCAÇÃO, ESCALA E FORMA

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Pernambuco como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dra. Fernanda De Bastiani

Recife-PE
2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Oliveira Júnior, Rondinelly Duarte de.
Seleção de Variáveis Explicativas nos Modelos Aditivos Generalizados de Locação, Escala e Forma / Rondinelly Duarte de Oliveira Júnior. - Recife, 2024.
45 p. : il., tab.

Orientador(a): Fernanda De Bastiani
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Estatística - Bacharelado, 2024.

1. Critério de informação de Akaike generalizado. 2. Critério de informação de Kullback-Leibler. 3. Regressão flexível. 4. Testes de Vuong e Clarke. I. De Bastiani, Fernanda . (Orientação). II. Título.

310 CDD (22.ed.)

Resumo

Este estudo visa explorar e comparar dois métodos de seleção de variáveis explicativas no contexto dos modelos aditivos generalizados de locação, escala e forma (GAMLSS). Um deles emprega o critério de informação de Akaike generalizado, enquanto o segundo realiza a seleção com base nos testes de Vuong e Clarke, estes sendo fundamentados na razão de verossimilhanças e utilizam o critério de informação de Kullback-Leibler. Para ambos os métodos, foi considerada a seleção de variáveis explicativas para a modelagem de todos os parâmetros da distribuição. Os métodos estão implementados no pacote **gamlss** da plataforma computacional `R`. Foram realizadas análise de dados proporcionando uma aplicação prática dos métodos estudados. A investigação se propõe a contribuir para a compreensão e adequação desses métodos de seleção de variáveis explicativas nos GAMLSS, oferecendo mecanismos para aprimorar a modelagem estatística em cenários complexos e diversos.

Palavras-chave: Critério de informação de Akaike generalizado; critério de informação de Kullback-Leibler; regressão flexível; testes de Vuong e Clarke.

Abstract

This study aims to explore and compare two methods of selection of explanatory variables in the context of the generalized additive models for location, scale and shape (GAMLSS). One of them uses the generalized Akaike information criterion, while the second performs the selection based on the Vuong and Clarke tests, which are based on the likelihood ratio and use the Kullback-Leibler Information Criterion. For both methods, the selection of explanatory variables was considered for modeling all distribution parameters. The methods were implemented in the **gamlss** package in `R` platform. Two real data analysis were carried out, providing a practical application of the methods studied. The investigation contributed to understand the adequacy of these methods for selecting explanatory variables in GAMLSS, offering new mechanisms to improve statistical modeling in complex and diverse scenarios.

Keywords: Flexible regression; generalized Akaike information criterion; Kullback-Leibler information criterion; Vuong and Clarke tests.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Nas últimas décadas, houve avanços consideráveis no aprimoramento de modelos estatísticos de regressão, com a criação de técnicas cada vez mais refinadas para lidar com a complexidade de conjuntos de dados. Essa área de crescente interesse e relevância envolve a integração de métodos estatísticos sofisticados, como os GAMLSS (Generalized Additive Models for Location, Scale and Shape), uma ferramenta flexível para analisar dados complexos, introduzida por RIGBY; STASINOPOULOS (2005). Essa abordagem possibilita o uso de diversas distribuições de probabilidade para a variável resposta, tanto discretas, contínuas como mistas, além de permitir a modelagem de todos os parâmetros da distribuição como função de variáveis explicativas. Além disso, o método incorpora relações lineares e não lineares entre os termos explicativos e os parâmetros de ditribuições por meio de funções de suavização, como *P-splines* (STASINOPOULOS et al., 2017).

Um grande diferencial dos GAMLSS reside na capacidade de ajustar modelos de regressão mesmo quando a distribuição da variável resposta não pertence à família exponencial. Isso significa que é possível lidar com uma variedade de distribuições contínuas e discretas, capturando nuances como assimetria e curtose variadas. A obra de RIGBY et al. (2019) apresenta uma compreensão mais aprofundada das propriedades das distribuições disponíveis no pacote **gamlss** da plataforma estatística R (R Core Team, 2022).

A seleção de variáveis é uma etapa crucial e meticulosa no processo de modelagem estatística. Na literatura, é notável que a maioria das técnicas de seleção de variáveis explicativas concentra-se em modelos convencionais. Contudo, a aplicação direta dessas abordagens em modelos com estruturas mais complexas pode se mostrar inviável, pois isso pode resultar em um ajuste inadequado. Essa limitação ressalta a necessidade de técnicas mais flexíveis, capazes de lidar com modelos menos convencionais.

Dentro desse contexto, uma das estratégias adotadas neste estudo envolve a aplicação do procedimento *stepwise*, obtido por meio da função `stepGAIC()` do pacote `gamlss`. Esse método dinâmico permite a eliminação ou adição de variáveis em cada fase do processo da modelagem, agindo sempre que essa alteração resulta em uma redução no critério (STASINOPOULOS et al., 2017), ou seja, a cada iteração do procedimento, o modelo é ajustado de acordo com as variáveis selecionadas, refinando continuamente a configuração do modelo com base na otimização do critério de informação de akaike generalizado (GAIC). Neste trabalho, a tarefa de selecionar termos explicativos para todos os parâmetros da distribuição será realizada pela função `stepGAICAll.A()`, também conhecida como Estratégia A, que oferece um sistema iterativo na identificação das variáveis que mais contribuem para a adequação do modelo aos dados.

Outra técnica de seleção de modelos utilizada nesse estudo que, embora menos mencionadas, apresentam-se com ótimo desempenho, é a função `VC.test()`, baseada

nos testes propostos por Vuong (VUONG, 1989) e Clarke (CLARKE, 2007). Esses testes, fundamentados na razão de verossimilhanças, utilizam o critério de informação de Kullback-Leibler (KULLBACK; LEIBLER, 1951), e revelam-se particularmente úteis quando se trata da escolha entre dois modelos que não necessariamente estejam aninhados. Essa característica faz deles uma alternativa eficaz e flexível, pois além de permitir a comparação entre modelos com termos diferentes, também possibilita comparar modelos com distribuições diferentes.

Este estudo visa evidenciar a importância dos testes de Vuong e Clarke na seleção de termos explicativos em modelos não aninhados. Por meio da análise comparativa dessas técnicas, busca-se demonstrar como esses testes podem desempenhar um papel na identificação e escolha adequada de variáveis explicativas nos GAMLSS. Para consolidar a seleção efetuada pelos testes, será conduzida uma análise dos resíduos por meio de ferramentas gráficas específicas, como o *worm plot* e o *bucket plot*, (BUUREN; FREDRIKS, 2001) e (BASTIANI et al., 2022). Esses gráficos são fundamentais para avaliar a adequação das distribuições e dos modelos de variáveis explicativas aos parâmetros da distribuição. Eles fornecem informações acerca da assimetria e curtose. Portanto, pretende-se com essa investigação contribuir para o conhecimento na área da regressão, proporcionando uma compreensão mais aprofundada sobre o desempenho desses testes, com potenciais implicações na melhoria das práticas de modelagem.

2 Revisão de Literatura

Nesta seção, foi exposta uma síntese detalhada das informações fundamentais obtidas por meio de levantamentos bibliográficos que abrange livros e artigos, e se concentra na metodologia dos GAMLSS. Ademais, foram exploradas algumas estratégias relacionadas à seleção de variáveis explicativas, bem como à análise de diagnósticos de resíduos.

Em se tratando de técnicas de modelagem de regressão univariada, destacam-se o modelo linear generalizado (em inglês, generalized linear models - GLM) (NELDER; WEDDERBURN, 1972) e o modelo aditivo generalizado (em inglês, generalized additive models - GAM) (HASTIE, 2017). Em ambos os modelos, a variável resposta, normalmente denotada por y, assume uma distribuição que pertence à família exponencial, em que a média μ de y é modelada como uma função de variáveis explicativas. A variância de y, expressa por $V(y) = \phi v(\mu)$, depende de um parâmetro de dispersão constante ϕ e da média μ, por meio da função de variância $v(\mu)$. Além disso, a assimetria e curtose de y não são explicitamente modeladas por variáveis explicativas, pois elas são geralmente funções de μ e ϕ. Portanto, nos modelos GLM e GAM, a variância, assimetria e curtose não são modeladas explicitamente em termos das variáveis explicativas, mas sim implicitamente por meio de sua dependência em relação a μ.

A partir do GLM e GAM, pôde-se obter diversas classes de modelos, como o Modelos Lineares Mistos (GLMMs) e os Modelos Mistos Aditivos Generalizados (GAMMs). Os GLMMs são extensões dos GLMs e assumem normalidade para a distribuição condicional de y dado os efeitos aleatórios (geralmente normais) (BRESLOW; CLAYTON, 1993). Enquanto os GAMMs são apenas GLMMs em que parte do preditor linear é especificada em termos de funções de suaviação de covariáveis (WOOD, 2017). Os GAMMs são propostos para dados muito dispersos e correlacionados, que surgem frequentemente em estudos envolvendo designs agrupados, hierárquicos e espaciais. Esta classe de modelos permite a dependência funcional flexível de uma variável de resultado em covariáveis usando regressão não paramétrica, ao mesmo tempo em que leva em conta a correlação entre observações usando efeitos aleatórios (LIN; ZHANG, 1999).

RIGBY; STASINOPOULOS (2005) introduziram em seu artigo uma classe geral de modelos de regressão univariada denominada Modelo Aditivo Generalizado de Localização, Escala e Forma (GAMLSS). Nessa abordagem, a suposição da família exponencial é substituída por uma família de distribuições bastante ampla. Dentro dessa nova estrutura, foi expandida a parte sistemática do modelo para permitir que não apenas a média μ (ou locação), mas todos os parâmetros da distribuição condicional de y, sejam modelados como funções aditivas (suaves) paramétricas e/ou não paramétricas das variáveis explicativas e/ou termos de efeitos aleatórios, assim possibilitando que a variância, assimetria e curtose sejam modeladas como função de variáveis explicativas. O

ajuste dos GAMLSS é alcançado por meio de dois procedimentos algorítmicos distintos, sendo o primeiro algoritmo baseado no método utilizado para ajustar os modelos aditivos de média e dispersão (RIGBY; STASINOPOULOS, 1996); e o segundo fundamentado no algoritmo desenvolvido por COLE; GREEN (1992).

RIGBY; STASINOPOULOS (2005) também apresentou o GAIC (AKAIKE, 1983) como uma extensão do AIC (AKAIKE, 1973) a fim de lidar com a complexidade associada aos GAMLSS, em que os modelos podem incluir múltiplos parâmetros para diferentes momentos e formas da distribuição condicional da variável resposta. Em essência, o GAIC ajusta o AIC para modelos GAMLSS, levando em consideração a dificuldade introduzida pela modelagem de diferentes parâmetros da distribuição.

RIGBY et al. (2019), em seu livro, abordou mais de 100 distribuições presentes no pacote `gamlss.dist` da plataforma R, explorando suas características, limitações e aplicabilidades em contextos de dados. Segundo os autores, historicamente, as distribuições normal e de Poisson eram comumente empregadas para modelar variáveis resposta contínua e discreta, respectivamente, no entanto, diante dos extensos conjuntos de dados complexos, essas distribuições são frequentemente consideradas inadequadas ou incapazes de fornecer informações adequadas.

No artigo de STASINOPOULOS; RIGBY (2008), são aparesentadas algumas funções do pacote **gamlss** para auxiliar na seleção de variáveis explicativas, incluindo as funções `addterm()` e `dropterm()`, que permitem, respectivamente, a adição ou remoção de um termo em um modelo. Essas duas funções são bases para muitas outras, como as funções baseadas no Critério da Informação de Akaike Generalizado. No livro de STASINOPOULOS et al. (2017) são apresentadas duas estratégias para selecionar termos para todos os parâmetros de uma distribuição, denominadas de Estratégia A e Estratégia B. Estas são implementadas usando as funções `stepGAICAll.A()` e `stepGAICAll.B()`, respectivamente. No artigo de STASINOPOULOS; RIGBY; BASTIANI (2018) os autores apresentam um tutorial de como selecionar modelos por meio da análise de dois conjuntos de dados.

RAMIRES et al. (2021), em seu artigo, realizaram vários estudos de simulação para investigar se uma abordagem específica baseada em *stepwise*, ou seja, a Estratégia A, seleciona adequadamente variáveis autênticas nos GAMLSS. Os autores consideraram as distribuições gaussianas, Poisson inflacionadas de zero e Weibull para a simulação. Além disso, foram consideradas variáveis explicativas contínuas (com relações lineares e não lineares) e categóricas. Os resultados do estudo foram satisfatórios, pois a Estratégia A mostrou um ótimo desempenho.

No artigo de RIGBY; STASINOPOULOS (2014), é apresentado um procedimento para a seleção automatizada dos parâmetros de suavização em um modelo GAMLSS. Os autores estabelecem um método que emprega uma representação *P-spline* dos termos de suavização, expressando-os como termos de efeito aleatório com uma estimativa interna (ou

local) de máxima verossimilhança na escala do preditor de cada parâmetro de distribuição, visando estimar seus parâmetros de suavização. A aplicação do método concentra-se na estimativa percentil, em que os quatro parâmetros de uma distribuição para a variável resposta são modelados como funções de suavização de uma variável explicativa transformada.

STASINOPOULOS et al. (2023) apresentam em seu artigo *P-splines* como uma ferramenta versátil de modelagem estatística, lidando com relações não lineares entre a resposta e variáveis explicativas. Os autores afirmam que a combinação da metodologia dos GAMLSS e *P-splines* fornece uma das ferramentas mais poderosas na análise de regressão moderna. O artigo discute a aplicação das duas técnicas quando a variável resposta é ajustada a zero (ou semicontínua).

O artigo de QU et al. (2020) aborda a análise de frequência de inundação em ambientes dinâmicos, concentrando-se no método não estacionário. Para superar limitações pela complexidade, o estudo propôs o modelo GAMLSS-CB, baseado no método de suavização *B-spline* Cúbico. O critério de seleção utilizado foi o GAIC.

RIGBY; STASINOPOULOS (2006), em sua pesquisa, apresentam a distribuição Box-Cox t (BCT) como um modelo para uma variável dependente. Os autores mencionam que os GAMLSS são expandidos para permitir que cada um dos parâmetros da distribuição seja modelado como funções paramétricas lineares e/ou não lineares e/ou funções de suavização não paramétricas das variáveis explicativas. Além disso, eles abordam um algoritmo de Fisher para ajustar o modelo e maximizar a verossimilhança (penalizada).

VONCKEN; ALBERS; TIMMERMAN (2019), em seu artigo, aplicaram um estudo de simulação, para investigar o desempenho de dois procedimentos de seleção de modelo *stepwise*, combinados com quatro critérios de ajuste de modelo (AIC, BIC, GAIC e validação cruzada), usando a distribuição BCPE (RIGBY; STASINOPOULOS, 2004). Os autores mostraram que o procedimento GAIC destacou-se como o método mais eficiente em termos de exigência de tamanho de amostra, sendo, portanto, o procedimento de seleção de modelo preferencial.

VUONG (1989) introduziu, em seu artigo, um teste clássico para seleção de modelos baseado no Critério de Informação de Kullback-Leibler. O teste é baseado em razões de verossimilhança para testar a hipótese nula de que os modelos concorrentes estão igualmente próximos do verdadeiro processo de geração de dados contra a hipótese alternativa de que um modelo está mais próximo. Os testes são direcionais e são derivados sucessivamente para os casos em que os modelos concorrentes são não aninhados, sobrepostos ou aninhados e se ambos, um ou nenhum estão especificados incorretamente.

SCHNEIDER et al. (2020) publicaram uma aplicação do teste de Vuong. No artigo, foi utilizada a seleção de modelos de Vuong para comparar modelos de Teoria de Resposta ao Item (TRI) unidimensionais e multidimensionais aninhados e não aninhados. Os autores argumentaram que a abordagem de Vuong fornece um conjunto útil de ferramentas e reforçaram a eficácia de utilizar esse método para comparar modelos de TRI aninhados e

não aninhados concorrentes.

CLARKE (2007) introduziu, em seu trabalho, um teste simples para seleção de modelos não aninhados. Segundo o autor, o teste mostrou-se assintoticamente mais eficiente do que o teste Vuong quando a distribuição das razões de log-verossimilhança individuais é altamente elevada. No estudo, a simulação demonstrou que o teste proposto teve maior potência que o teste Vuong na aplicação ao efeito das instituições políticas nacionais na tomada de decisões em política externa.

DIXIT; JAYAKUMAR (2022), abordaram os GAMLSS e os testes de Vuong e Clarke, em seu estudo, para dados relacionados à caracterização de seca usando modelo de construção de cópula trivariada e pareada. As famílias de cópulas para as variáveis de pares foram selecionadas de várias cópulas com base nos testes de Clarke e Vuong.

BUUREN; FREDRIKS (2001) apresentaram em seu artigo uma ferramenta de diagnóstico, por meio de visualização gráfica, chamado de *Worm Plot*, pra análise de resíduos. De acordo com o autor, o *Worm Plot* avalia a normalidade dos resíduos, bem como assimetria e curtose, de forma satisfatória.

BASTIANI et al. (2018) utilizaram o *Worm Plot* na análise de resíduos, publicado em um artigo que descreve a modelagem e ajuste de componentes espaciais de campo aleatório Gaussiano de Markov no contexto dos GAMLSS, que permite a modelagem de todos os parâmetros da distribuição para a variável resposta usando variáveis explicativas com efeitos espaciais.

No artigo de HOSSAIN et al. (2016) *Worm Plots* foram utilizados como um método para verificar a adequação de 4 modelos ajustados diferentes que estimam curvas percentuais, inclusive um dos modelos utiliza a distribuição BCCGo (Box-Cox Cole Green orig.). Os critérios de seleção utilizados foram do AIC e BIC.

BASTIANI et al. (2022) introduziram, em seu estudo, um outro método de visualização chamado de *Bucket Plot*, que se trata de uma ferramenta visual para detectar assimetria e curtose em uma variável resposta ou nos resíduos de um modelo ajustado, assim sendo eficaz para avaliar a adequação de uma distribuição ajustada à variável resposta. No artigo, foi demonstrado o *Bucket Plot* em nove cenários simulação de assimetria e curtose.

No artigo de STASINOPOULOS; RIGBY (2008), é definida a estrutura estatística do GAMLSS e descrita a implementação do GAMLSS na plataforma computacional R. Além disso, os autores fornecem quatro exemplos de dados diferentes para demonstrar como o GAMLSS pode ser usado para modelagem estatística.

Dessa forma, esse levantamento bibliográfico tem o intuito de destacar o contexto da pesquisa no qual a proposta em questão está inserida, proporcionando uma visão das principais contribuições. Ao examinar as realizações dos últimos anos, espera-se que essa revisão possa auxiliar este estudo de forma a gerar um trabalho que contribua para o avanço contínuo nesse campo de estudo que engloba os modelos GAMLSS.

3 Objetivos

3.1 Objetivo geral

O objetivo deste estudo é explorar métodos de seleção de variáveis explicativas no contexto dos modelos aditivos generalizados de locação, escala e forma. Em particular, explorar o critério de Informação de Akaike Generalizado (GAIC) e os testes de Vuong e Clarke para a seleção de variáveis explicativas nos GAMLSS.

3.2 Objetivos específicos

- Analisar detalhadamente a aplicação do GAIC na seleção de variáveis explicativas nos GAMLSS;
- Investigar os fundamentos teóricos e práticos dos testes de Vuong e Clarke para a seleção de variáveis explicativas nos GAMLSS;
- Aplicar os métodos à análise de dados reais;
- Fornecer orientações para a escolha do modelo final.

4 Metodologia

4.1 GAMLSS

Os GAMLSS oferecem um método muito vasto e flexível para modelar uma variável resposta. Na modelagem, é possível selecionar a distribuição da variável resposta em meio às inúmeras distribuições disponíveis no pacote **gamlss** da plataforma computacional R, que podem ser distribuições contínuas, discretas ou mistas, com alto grau de assimetria e curtose. As distribuições disponíveis nesse pacote podem ter até quatro parâmetros, denotados por (μ, σ, ν, τ), em que os dois primeiros parâmetros (μ, σ) são caracterizados como parâmetros de locação e escala, enquanto os parâmetros (ν, τ) modelam a forma da distribuição, isto é, a assimetria e a curtose. Em (RIGBY; STASINOPOULOS, 2005), o modelo GAMLSS é definido como apresentado a seguir.

Seja $\boldsymbol{y}^T = (y_1, y_2, \ldots, y_n)$ o vetor das observações da variável resposta. Seja $g_k(\cdot)$ uma função de ligação monótona conhecida, para $k = 1, 2, \ldots, p$, relacionando θ_k às variáveis explicativas e aos efeitos aleatórios por meio de um modelo aditivo dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\gamma_{jk} \tag{4.1}$$

em que $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores de comprimento n, por exemplo, $\boldsymbol{\theta}_k^T = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kn})$; $\boldsymbol{\beta}_k^T = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{J'_k k})$ é um vetor de parâmetros de comprimento J'_k; $\boldsymbol{X}_k$, é uma matriz conhecida de ordem $n \times J'_k$; $\boldsymbol{Z}_{jk}$ é uma matriz fixa conhecida $n \times q_{jk}$; e γ_{jk} é uma variável aleatória q_{jk}-dimensional. Então, a Equação (4.1) é chamada de GAMLSS.

Se $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, em que $\boldsymbol{I}_n$ é uma matriz identidade $n \times n$, e $\gamma_{jk} = \boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ para todas combinações de j e k no modelo da Equação (4.1), tem-se

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\boldsymbol{x}_{jk}) \tag{4.2}$$

em que $\boldsymbol{x}_{jk}$ são vetores de comprimento n, para $j = 1, 2, \ldots, J_k$ e $k = 1, 2, \ldots, p$. A função h_{jk} é uma função desconhecida da variável explicativa X_{jk}, e $\boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ é o vetor que avalia a função h_{jk} em $\boldsymbol{x}_{jk}$. Os vetores explicativos $\mathbf{x}_{jk}$ são assumidos como conhecidos. O modelo na Equação (4.2) é chamado de modelo GAMLSS semiparamétrico.

Para muitas famílias de distribuições populacionais, um máximo de dois parâmetros de forma $\nu = \theta_3$ e $\tau = \theta_4$ é suficiente, têm-se

$$
\begin{aligned}
g_1(\mu) = \eta_1 &= \boldsymbol{X}_1\beta_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\gamma_{j1}, \\
g_2(\sigma) = \eta_2 &= \boldsymbol{X}_2\beta_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\gamma_{j2}, \\
g_3(\nu) = \eta_3 &= \boldsymbol{X}_3\beta_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\gamma_{j3}, \\
g_4(\tau) = \eta_4 &= \boldsymbol{X}_4\beta_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\gamma_{j4}.
\end{aligned}
\tag{4.3}
$$

Os GAMLSS oferecem muitas vantagens na modelagem de dados, por exemplo, todos os parâmetros da distribuição da variável resposta podem ser modelados utilizando funções de suavização, paramétricas e/ou não paramétricas, de variáveis explicativas. Além disso, diferentes distribuições que não fazem parte da família exponencial podem ser ajustadas.

4.2 Funções do Pacote **gamlss**

As funções básicas para seleção de variáveis explicativas no pacote **gamlss** são `addterm()` e `dropterm()`, que permitem adicionar ou remover um termo para um parâmetro de distribuição, respectivamente. Essas funções são procedimentos para `stepGAIC()`, que é adequada para a seleção de variáveis, usando o procedimento *stepwise*. `stepGAIC()` é, por sua vez, o procedimento básico para `stepGAICAll.A()` e `stepGAICAll.B()` (STASINOPOULOS et al., 2017).

A função `stepGAIC()` pode ser usada para construir um modelo para qualquer um dos parâmetros de distribuição usando um procedimento *foward*, *backward* ou *stepwise* (a busca *stepwise* pode ser tanto *backward* ou *forward*), baseado no Critério dE Informação de Akaike Generalizado. Ela é baseado na função `stepAIC()` ((RIPLEY, 2002)). Adicionalmente, se o argumento de escopo especificado no código estiver faltando, o padrão será seguir a direção *backward*.

4.2.1 Funções `fitDist()` e `ChooseDist()`

Para a escolha da distribuição a serem implementadas nos ajustes deste estudo, foram utilizadas as funções `fitDist()` e `ChooseDist()`.

A função `fitDist()` emprega o uso da função `gamlssML()` para adaptar todas as distribuições paramétricas pertinentes, especificadas pelo argumento "`type`" da família `gamlss.family` a um único conjunto de dados, o qual não inclui variáveis explicativas. A distribuição marginal final é determinada pelo critério de informação generalizada de Akaike com penalização `k`. O valor padrão para `k` é `2`, correspondendo ao AIC.

A função `ChooseDist()` utiliza a função `update.gamlss()` para ajustar todas as distribuições paramétricas relevantes da família `gamlss.family` a um modelo GAMLSS específico já ajustado. Além disso, a função leva em consideração as variáveis explicativas do modelo. A saída da função é uma matriz que contém diferentes distribuições nas linhas e os diferentes valores de `GAIC()` nas colunas. O valor padrão para o parâmetro `k` é `2`, para AIC, e `log(n)` para BIC.

4.2.2 Estratégia A: `stepGAICAll.A()`

Como mencionado, a função `stepGAICAll.A()` é baseada na função `stepGAIC()`. A Estratégia A é utilizada para selecionar termos aditivos, usando o GAIC, para todos os parâmetros de uma distribuição, assumindo uma distribuição particular. O objetivo da estratégia é encontrar modelos apropriados para os diferentes parâmetros da distribuição.

O modelo final possivelmente terá diferentes subconjuntos de termos para μ, σ, ν e τ. Em (STASINOPOULOS et al., 2017), são apresentados os passos abaixo para aplicação da Estratégia A:

- i) construir um modelo para μ usando uma abordagem direta;
- ii) dado o modelo para μ, construir um modelo para σ (*forward*);
- iii) dados os modelos para μ e σ, construir um modelo para ν (*forward*);
- iv) dados os modelos para μ, σ e ν, construir um modelo para τ (*forward*);
- v) dados os modelos para μ, σ, ν e τ, verifique se os termos para ν são necessários usando eliminação *backward*;
- vi) dados os modelos para μ, σ, ν e τ, verifique se os termos para σ são necessários (*backward*);
- vii) dados os modelos para μ, σ, ν e τ, verifique se os termos para μ são necessários (*backward*).

4.2.3 Função `VC.test()`

A função `VC.test()` é usada para selecionar modelos estatísticos por meio dos testes de Vuong e Clarke e tem o seguinte formato no R:

```
VC.test(obj1, obj2, sig.lev = 0.05)
```

em que

- `obj1`: o primeiro modelo GAMLSS ajustado;

- `obj2`: o segundo modelo GAMLSS ajustado;
- `sig.lev`: nível de significância usado para o teste.

Esses testes são baseados em razões de verossimilhança e têm como objetivo a seleção de modelos utilizando o critério de informação de Kullback-Leibler (KLIC) (KULLBACK; LEIBLER, 1951). Mais ainda, os testes são aplicados para a comparação de dois modelos bivariados, que não são necessariamente aninhados.

Vuong define o KLIC como

$$KLIC \equiv E_0[\ln h_0(Y_i|X_i)] - E_0[\ln f(Y_i|X_i;\beta_*)], \tag{4.4}$$

em que $h_0(.|.)$ é a verdadeira densidade condicional de Y_i dado X_i, (isto é, o modelo verdadeiro, mas desconhecido), E_0 é a esperança sob o modelo verdadeiro, e β_* são os valores pseudo-verdadeiros de β (as estimativas de β quando $f[Y_i|X_i]$ não é o modelo verdadeiro). O melhor modelo é aquele que minimiza a Equação 4.4, pois o melhor modelo é aquele que mais se aproxima da especificação verdadeira. Deve-se, portanto, escolher o modelo que maximiza $E_0[\ln f(Y_i|X_i;\beta_*)]$. Em outras palavras, um modelo deve ser selecionado em detrimento de outro se a probabilidade logarítmica média desse modelo for significativamente maior do que a probabilidade logarítmica média do modelo rival.

4.2.4 Vuong Test

No teste de Vuong, a hipótese nula é que os modelos concorrentes estão igualmente próximos do verdadeiro processo de geração de dados, contra a hipótese alternativa de que um modelo está mais próximo. Sob as condições gerais de que os modelos podem ser aninhados, não aninhados ou sobrepostos, e que ambos, apenas um ou nenhum dos modelos concorrentes podem conter a verdadeira lei que gera as observações. O teste considera o KLIC, que mede a distância entre uma determinada distribuição e a verdadeira distribuição. Ou seja, se for considerada a distância entre um modelo específico e a distribuição verdadeira como o mínimo do KLIC entre as distribuições no modelo, é natural definir o melhor modelo em um conjunto de modelos concorrentes como aquele que está mais próximo da verdadeira distribuição. Além disso, o teste segue assintoticamente uma distribuição normal padrão sob o valor nulo.

Agora, considere os modelos condicionais para incorporar variáveis explicativas. Então, se $F_\theta = \{f(y|z;\theta;\theta \in \Theta\}$ é um modelo condicional, a distância entre sua densidade condicional verdadeira $h^0(y|z)$, medida pelo KLIC mínimo, é representada por $E^0[\log h^0(y|z)] - E^0[\log f(y|z;\theta_*)]$, onde $E^0[.]$ denota a esperança em relação à verdadeira distribuição conjunta de (y,z) e θ_* é o valor pseudo-verdadeiro de θ. Então, um critério de seleção equivalente pode ser formulado com base na quantidade $E^0[\log f(y|z;\theta_*)]$, sendo o melhor modelo aquele em que essa quantidade é maximizada (VUONG, 1989).

4.2.4.1 Hipóteses do Teste

Considere dois modelos, $\boldsymbol{F}_\beta = f(Y_i|X_i;\beta)$ e $\boldsymbol{G}_\gamma = g(Y_i|Z_i;\gamma)$. A hipótese nula do teste é dada na Equação 4.5.

$$H_0 : E_0\left[\ln\frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)}\right] = 0, \tag{4.5}$$

o que indica que dois modelos rivais estão igualmente próximos da especificação verdadeira. Vuong prova sob condições gerais que o valor esperado dado na hipótese nula pode ser consistentemente estimado por $(1/n)$ vezes a estatística da razão de verossimilhança, como mostra a Equação 4.6.

$$\frac{1}{n}LR_n(\hat{\beta}_n,\hat{\gamma}_n) \xrightarrow{a.s.} E_0\left[\ln\frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)}\right], \tag{4.6}$$

em que $\hat{\beta}_n$ e $\hat{\gamma}_n$ são os estimadores de máxima verossimilhança de β_* e γ_*. A estatística de razão de verossimilhança resultante é assintoticamente normalmente distribuída e o teste real é, portanto,

$$\text{sob } H_0 : \frac{LR_n(\hat{\beta}_n,\hat{\gamma}_n)}{(\sqrt{n})\hat{\omega}_n} \xrightarrow{D} N(0,1), \tag{4.7}$$

em que o numerador é a diferença nas log-verossimilhanças somadas para os dois modelos, $LR_n(\hat{\beta}_n,\hat{\gamma}_n) \equiv L_n^f(\hat{\beta}_n) - L_n^g(\hat{\gamma}_n)$, e $\hat{\omega}_n$ é o desvio padrão estimado calculado da maneira usual,

$$\hat{\omega}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{f(Y_i|X_i;\hat{\beta}_n)}{g(Y_i|Z_i;\hat{\gamma}_n)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(Y_i|X_i;\hat{\beta}_n)}{g(Y_i|Z_i;\hat{\gamma}_n)}\right]^2 \tag{4.8}$$

A estatística Vuong é sensível ao número de coeficientes estimados em cada modelo e, portanto, o teste deve ser corrigido para a dimensionalidade do modelo. Vuong (1989) sugere o uso de uma correção que corresponda aos critérios de informação de (AKAIKE, 1973) ou aos critérios de informação bayesianos de (SCHWARZ, 1978). Usando este último, a estatística ajustada torna-se

$$L\tilde{R}_n(\hat{\beta}_n,\hat{\gamma}_n) - \left[\left(\frac{p}{2}\right)\ln n - \left(\frac{q}{2}\right)\ln n\right], \tag{4.9}$$

em que p e q são o número de coeficientes estimados nos modelos $\boldsymbol{F}_\beta$ e $\boldsymbol{G}_\gamma$, respectivamente.

Para o resultado do teste, considere uma região crítica definida como $(-c, c)$, onde c geralmente assume o valor de 1,96. Se o valor do teste for superior a c, rejeitamos a hipótese nula de que os modelos são equivalentes, favorecendo assim o modelo no `obj1`. Por outro lado, se o valor for inferior a $-c$, rejeitamos a hipótese nula em favor do modelo no `obj2`. Quando o valor está dentro do intervalo $(-c, c)$, não é possível discriminar os dois modelos concorrentes.

4.2.4.2 Clarke Test

O teste de Clarke mostra-se assintoticamente mais eficiente do que o teste Vuong quando a distribuição das razões de log-verossimilhança individuais é altamente elevada (com relação à distribuição normal).

4.2.4.2.1 Hipóteses do Teste

A alternativa livre de distribuição de (CLARKE, 2007) aplica um teste de sinal pareado modificado às diferenças nas log-verossimilhanças individuais de dois modelos não aninhados. Usando a notação de Vuong, a hipótese nula do teste livre de distribuição é

$$H_0 : Pr_0\left[\ln \frac{f(Y_i|X_i;\beta_*)}{g(Y_i|Z_i;\gamma_*)} > 0\right] = 0,5. \tag{4.10}$$

A Equação 4.10 afirma que, sob a hipótese nula, as razões log-verossimilhança devem ser distribuídas uniformemente em torno de zero. Assim, metade das razões log-verossimilhança deve ser maior que zero e metade menor que zero. A diferença entre as Equações 4.10 e 4.5 é que a esperança na Equação 4.5 é substituída pela mediana na Equação 4.10.

Seja $d_i = \ln f(Y_i|X_i;\hat{\beta}_n) - \ln g(Y_i|Z_i;\hat{\gamma}_n)$. A estatística de teste é dada pela Equação 4.11.

$$B = \sum_{i=1}^{n} I_{(0,+\infty)}(d_i), \tag{4.11}$$

em que I é a função do indicadora. A Equação 4.11 é o número de diferenças positivas e é distribuída pela Binomial com parâmetros n e $\theta = 0,5$.

Se o modelo $\boldsymbol{F}_\beta$ for melhor que o modelo $\boldsymbol{G}_\gamma$, B será significativamente maior que seu valor esperado sob a hipótese nula $(n/2)$. Para um teste de cauda superior, rejeitamos a hipótese nula de equivalência quando $B \geq c_\alpha$, onde c_α é escolhido como o menor inteiro tal que $\sum_{c=c_\alpha}^{n} \binom{n}{c} 0,5^n \leq \alpha$.

O teste de Clarke, assim como o teste de Vuong, é sensível à dimensionalidade dos modelos concorrentes. Uma vez que estamos lidando com as razões de log-verossimilhança individuais, não é possível aplicar a mesma correção à razão de log-verossimilhança somada, como feito por Vuong em seu teste. Contudo, podemos empregar a correção média nas razões de log-verossimilhança individuais. Em outras palavras, ajustamos as log-verossimilhanças individuais para o modelo $\boldsymbol{F}_\beta$ por um fator $[(p/2n)\ln n]$ e as log-verossimilhanças individuais para o modelo $\boldsymbol{G}_\gamma$ por um fator $[(q/2n)\ln n]$.

Intuitivamente, o modelo no `obj1` é preferível ao de `obj2` se o valor do teste for significativamente maior que seu valor esperado sob a hipótese nula e vice-versa. Se o valor não for significativamente diferente de $n/2$, então `obj1` pode ser considerado equivalente a `obj2`.

4.2.5 Critério da Informação de Akaike Generalizado (GAIC)

O Critério de Informação de Akaike Generalizado (GAIC, do inglês Generalized Akaike Information Criterion) é um método utilizado na seleção de modelos estatísticos. O GAIC é uma extensão do Critério de Informação de Akaike (AIC) e é projetado para modelos GAMLSS, que são utilizados para modelar diferentes parâmetros da distribuição de uma variável resposta. O GAIC leva em consideração a verossimilhança do modelo, o número de parâmetros estimados e o tamanho da amostra para penalizar modelos mais complexos. O GAIC é definido como

$$GAIC(\kappa) = -2\hat{l}(\hat{\boldsymbol{\theta}}) + \kappa \times df \tag{4.12}$$

em que *df* representa os graus de liberdades efetivos usados para o modelo ajustado; κ é a penalização para cada grau de liberdade usado; $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\tau}})$ são os parâmetros ajustados, os quais dependem das variáveis explicativas; $\hat{l}(\hat{\boldsymbol{\theta}})$ é a log-verossimilhança ajustada e $-2\hat{l}(\hat{\boldsymbol{\theta}})$ é o desvio global, sendo

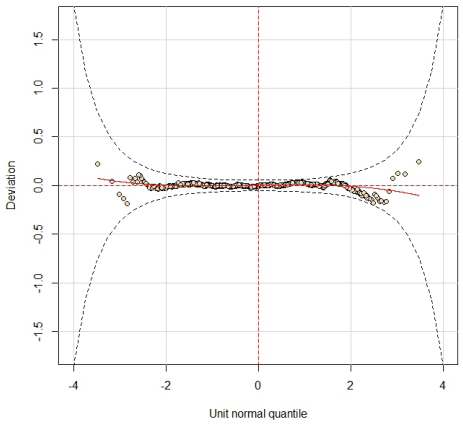
$$\text{Desvio Global} = -2\hat{l}(\hat{\boldsymbol{\theta}}) = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) \tag{4.13}$$

Para mais informações acerca do desvio global, consulte a Seção 10.2 em (RIGBY et al., 2019).

4.2.6 *Worm Plot*

Worm plot (gráfico de minhoca, em portugês) é um tipo de gráfico QQ-plot sem tendência utilizado nos diagnósticos para análise de resíduos nos GAMLSS. O termo "minhoca" refere-se à aparência de minhoca dos pontos plotados no gráfico. Essa ferramenta de visualização foi apresentada por van Buuren e Fredriks (BUUREN; FREDRIKS, 2001) com o objetivo de identificar intervalos de uma variável explicativa em que o modelo não se ajusta adequadamente aos dados. No R, o gráfico é implementado por meio da função `wp()`, que gera gráficos individuais ou múltiplos do tipo *worm plot* para modelos ajustados nos GAMLSS. Essa função é utilizada para examinar os resíduos em diferentes intervalos de uma ou duas variáveis explicativas. Como ilustração, utilizaremos a Figura 1, que se refere ao modelo denominado `M_bccgo_suav_k2`, ajustado na Seção 5.3.2.2.

Figura 1 – Ilustração de *worm plot*.



Fonte: Figura 7a da Seção 5.3.2.2.

Os pontos no gráfico indicam a distância entre os resíduos ordenados e seus valores esperados, os quais são delineados pela linha horizontal pontilhada na Figura 5.3.2.2. Espera-se, portanto, que os pontos plotados estejam próximos à linha horizontal, pois quanto mais próximos os pontos estiverem dessa linha, maior será a conformidade da distribuição dos resíduos com uma distribuição normal padrão. Além disso, se o modelo estiver correto, espera-se que 95% dos pontos estejam entre as suas curvas elípticas da figura, pois elas representam os intervalos de confiança aproximados de 95%. Caso isso não aconteça, o gráfico indicará que a distribuição ajustada do modelo são inadequados para explicar a variável resposta. E ainda, é possível identificar valores discrepantes, quando os pontos estão muito distantes das curvas elípticas (STASINOPOULOS et al., 2017).

A curva que passa pelos pontos é um ajuste cúbico aos pontos do gráfico. A forma deste ajuste cúbico pode refletir diferentes inadequações no modelo, descritas na Tabela 1.

A Tabela 1 mostra algumas preocupações em relação ao ajuste. Como o caso dos pontos acima da linha horizontal, indicando que a média residual é muito alta, o que implica que a locação da distribuição ajustada é muito baixa. Isso pode ser corrigido aumentando o parâmetro de locação μ, ou melhorando o modelo para μ, ou ainda mudando a distribuição do modelo. O mesmo pode ser feito quando o nível do gráfico estiver abaixo de uma linha horizontal na origem. Além disso, uma tendência linear (positiva ou negativa) indica problemas com a variância; a forma quadrática (U ou U invertido) aponta assimetria dos resíduos; e a forma cúbica (S com esquerda curvada para cima ou para baixo) indica curtose dos resíduos.

Tabela 1 – Interpretação da curva ajustada aos pontos em um *worm plot*.

Forma do *Worm Plot* (ou da curva ajustada)	**Resíduos**	**Distribuição ajustada**
Nível: acima da origem	média muito alta	locação ajustada muito baixa
Nível: abaixo da origem	média muito baixa	locação ajustada muito alta
Linha: inclinação positiva	variância muito alta	escala ajustada muito baixa
Linha: inclinação negativa	variância muito baixa	escala ajustada muito alta
Forma de U	assimetria positiva	assimetria ajustada muito baixa
Forma de U invertido	assimetria negativa	assimetria ajustada muito alta
Forma de S com a esquerda curvada para baixo	leptocúrtico	caudas da distribuição ajustada muito leves
Forma de S com a esquerda curvada para cima	platicúrtico	caudas da distribuição ajustada muito pesadas

Fonte: Adaptação da Tabela 12.1 do livro de STASINOPOULOS et al. (2017).

4.2.7 *Bucket Plot*

Bucket plot (gráfico balde, em portugês) é uma representação gráfica projetada para auxiliar na identificação da distribuição que melhor se adequa à variável resposta, especialmente em relação à assimetria e curtose (BASTIANI et al., 2022). O nome "*bucket plot*" é derivado de sua aparência visual, que se assemelha a um balde. Quando aplicado em uma análise de regressão ajustada, esse gráfico pode servir como uma ferramenta de diagnóstico, utilizando resíduos para avaliar a adequação da distribuição e dos modelos das variáveis explicativas em relação aos seus parâmetros. Se o modelo estiver correto, então os valores verdadeiros desses resíduos têm uma distribuição normal padrão (mesmo quando a distribuição do modelo não é normal).

Dentro de um *bucket plot*, a assimetria percentil e a curtose da variável de resposta (ou resíduos de um modelo ajustado para essa variável) são representadas como pontos em um gráfico de momento ou um gráfico de assimetria-curtose percentil. É importante observar que *bucket plots* têm a capacidade de identificar a presença de assimetria e/ou curtose, contudo, não são destinados a avaliar se os parâmetros de localização e escala estão sendo modelados de maneira apropriada.

De acordo com BASTIANI et al. (2022), o *bucket plot* de momento consiste nos seguintes componentes:

- Uma região em forma de balde, que representa o excesso de curtose do momento transformado permitido e a região de assimetria do momento transformada de todas as distribuições possíveis.

- A posição da amostra dentro do balde indica se há excesso de curtose, indicando se a distribuição é leptocúrtica (metade superior) ou platicúrtica (metade inferior), bem como a presença de assimetria negativa (metade esquerda) ou positiva (metade direita) em relação ao momento. Vale ressaltar que o ponto $(0, 0)$ representa os valores de uma distribuição normal. Amostras que se encontram próximas de $(0, 0)$,

indicando assimetria e excesso de curtose zero, são consistentes com a distribuição normal.

- Uma área elíptica sombreada ao redor do ponto $(0, 0)$ representa uma região de 95% de confiança para a assimetria e curtose, conforme avaliado pelo teste de Jarque-Bera (JARQUE; BERA, 1980). Se um ponto da amostra estiver dentro dessa região sombreada, não há evidência para rejeitar a hipótese nula do teste Jarque-Bera, indicando que a assimetria e a curtose são simultaneamente iguais a zero. No entanto, se o ponto estiver fora dessa região, a hipótese nula é rejeitada.

- Um conjunto de pontos que forma uma nuvem ao redor de um ponto central representa as medidas transformadas de assimetria e excesso de curtose da amostra. Cada ponto na nuvem corresponde a uma medida transformada de assimetria e excesso de curtose obtida a partir de amostras não paramétricas de *bootstrap* da variável. Cada uma dessas amostras de *bootstrap* possui o mesmo tamanho que a variável original. A disposição dos pontos na nuvem oferece uma representação visual da variabilidade associada à estimativa da assimetria do momento transformado e do excesso de curtose do momento.

4.2.8 *Term Plot*

O *term plot* (gráfico de termos, em português) é uma ferramenta gráfica utilizada em análises de regressão para visualizar a relação entre os termos de regressão e seus preditores específicos em um modelo ajustado. Esse método refere-se a gráficos que mostram o efeito de cada termo suavizado no modelo, sendo úteis para visualizar como cada variável de entrada contribui para a variável resposta do modelo (STASINOPOULOS et al., 2017).

A função `term.plot()` do pacote **gamlss**, é baseada na função `termplot()` padrão do `R`, e foi adaptada para ser aplicada as GAMLSS, permitindo representar graficamente os termos de regressão em relação aos seus preditores.

Cada termo pode ser examinado individualmente em relação aos seus preditores. O gráfico permite que você escolha um parâmetro específico do modelo GAMLSS para verificar a relação com seus preditores. Isso facilita a compreensão do efeito de cada variável em diferentes partes da distribuição condicional. Além disso, pode ser incluída informações sobre erros padrão e resíduos parciais na função. Essas informações adicionais ajudam na avaliação da qualidade do ajuste do modelo e na identificação de padrões não capturados pelos termos de regressão, como padrões não lineares e tendências.

O *term plot* é uma ferramenta importante principalmente para a interpretação dos resultados e pode servir também para diagnóstico, pois permite identificar possíveis problemas com o ajuste do modelo e áreas em que o modelo pode ser aprimorado. Ele proporciona uma visualização das relações funcionais entre variáveis e pode auxiliar na análise e no refinamento do modelo ajustado.

5 Resultados e Discussão

5.1 Base de Dados

5.1.1 Dados `rent`

A base de dados `rent`, disponível no R, é referente à uma pesquisa realizada em abril de 1993 pela Infratest Sozialforchung, na cidade de Munique, na Alemanha, e conta com 1969 observações para cada variável. As variáveis utilizadas para modelagem foram:

- `R`: representa o valor do aluguel mensal de apartamentos na região estudada;
- `Fl`: mostra o tamanho do apartamento em metros quadrados;
- `A`: caracteriza o ano de construção do imóvel;
- `B`: um fator com níveis indicando se existe um banheiro premium, 0, ou não, 1;
- `H`: indica se o imóvel tem aquecimento central, 0, ou não, 1;
- `loc`: representa a qualidade da localização, variando entre abaixo da média, 1, na média, 2, e acima da média, 3.

5.1.2 Dados `wines`

O conjunto de dados `wines` é composto por 13090 observações. A base contém cinco variáveis, cada uma representando uma característica específica dos vinhos. Todas foram consideradas relevantes para entender a distribuição de preços dos vinhos, são elas: País de origem (`Fc`), Avaliação média (`Rating`), Preço da garrafa (`Price`), Ano de produção (`Year`) e Tipo de vinho (`Ft`), sendo a `Price` a variável resposta do estudo. A base foi obtida no site *kaggle.com*.

5.1.2.1 Agrupamento da Variável `Fc`

Para a variável `Fc`, foi feito um agrupamento, pois há 33 países na base de dados, mas usar todos como *factor* poderia dificultar a análise. Com isso, o agrupamento foi feito com base em suas tradições vinícolas e reputação na produção de vinho e foram separados da seguinte forma:

Tradição Vinícola Forte (VinForte ou 1): Este grupo incluiu países que são amplamente reconhecidos por suas tradições vinícolas antigas, produção de vinhos de alta qualidade e fama internacional. França, Itália, Espanha e Portugal são considerados algumas das nações mais tradicionais na produção de vinho no mundo, com vinhedos

centenários, técnicas de vinificação tradicionais e uma longa história de exportação de vinhos de qualidade.

Tradição Vinícola Crescente (VinCrescente ou 2): Este grupo engloba países da América do Sul, Argentina e Chile, que ganharam destaque nas últimas décadas por sua produção de vinhos de qualidade crescente. Ambos os países têm vinhedos em altitude, climas adequados e variedades de uvas únicas que têm atraído a atenção do mercado global de vinho.

Tradição Vinícola Emergente (VinEmergente ou 3): Neste grupo, foram incluídos países como Estados Unidos, Austrália e Nova Zelândia, que emergiram como produtores de vinho de alta qualidade em um curto período de tempo. Eles têm adotado técnicas de vinificação inovadoras e são conhecidos por seus vinhos distintos.

Tradição Vinícola na Europa Central (VinEuropaCentral ou 4): Este grupo reúne países da Europa Central, como Alemanha, Áustria e Hungria, que têm uma rica tradição vinícola e são conhecidos por seus vinhos de alta qualidade, especialmente os brancos.

Tradição Vinícola no Leste Europeu (VinLesteEuropeu ou 5): Países do leste europeu, como Romênia, Bulgária, Moldávia e Geórgia, têm tradições vinícolas únicas e têm sido historicamente produtores de vinho.

Tradição Vinícola na América do Sul (VinAmericaSul ou 6): Além da Argentina e do Chile, o Brasil e o Uruguai também fazem parte deste grupo, representando países sul-americanos em ascensão na produção de vinho.

Tradição Vinícola na África do Sul (VinAfricaSul ou 7): A África do Sul é conhecida por sua tradição vinícola única, variedades de uvas distintas e sua história complexa na produção de vinho.

Tradição Vinícola na Grécia e no Oriente Médio (VinGreciaOriente ou 8): Este grupo reúne países como Grécia, Israel e Líbano, que têm tradições vinícolas antigas e estão começando a ganhar destaque no cenário internacional de vinhos.

Outros Países Produtores (Outros ou 9): Este grupo inclui países que não se encaixam facilmente em nenhuma das categorias anteriores, como Canadá, China, Turquia, México, Reino Unido e Luxemburgo. Eles podem ter tradições vinícolas menos reconhecidas ou uma produção vinícola menor em comparação com os outros grupos.

É importante observar que essas divisões são uma simplificação e que a indústria de vinho é complexa, com muitos países produzindo vinhos de alta qualidade em várias regiões. A divisão em grupos foi feita com base em critérios gerais, e há sempre exceções na produção de vinho em cada país. Segue o agrupamento:

Tabela 2 – Tradições Vinícolas.

VinForte:	**VinCrescente:**	**VinEmergente:**
França	Argentina	Estados Unidos
Itália	Chile	Austrália
Espanha		Nova Zelândia
Portugal		
VinEuropaCentral:	**VinLesteEuropeu:**	**VinAmericaSul:**
Alemanha	Romênia	Brasil
Áustria	Bulgária	Uruguai
Suíça	Moldávia	
Hungria	Geórgia	
República Tcheca		
Eslováquia		
Eslovênia		
Croácia		
`VinAfricaSul`:	**VinGreciaOriente:**	**Outros:**
África do Sul	Grécia	Canadá
	Israel	Suécia
	Líbano	China
		Turquia
		México
		Reino Unido
		Luxemburgo

5.2 Análise Descritiva da Base `rent`

Nesta seção, foi feita uma análise exploratória das variáveis utilizadas para modelar o aluguel mensal do conjunto de dados `rent`. A descrição desses dados pode ser encontrada na Seção 5.1.1. Aqui, deseja-se observar o comportamento dos dados e obter informações relevantes para o ajuste do modelo, como identificar possíveis distribuições candidatas para testes futuros.

Na Tabela 3, estão apresentadas algumas estatísticas descritivas da variável `R`, que representa o aluguel mensal em DM. Pode-se observar que o menor valor do aluguel de apartamento naquela região foi de 101,7 DM, enquanto o maior foi de 3000 DM, com um valor médio de 811,9 DM e um desvio padrão de 379 DM.

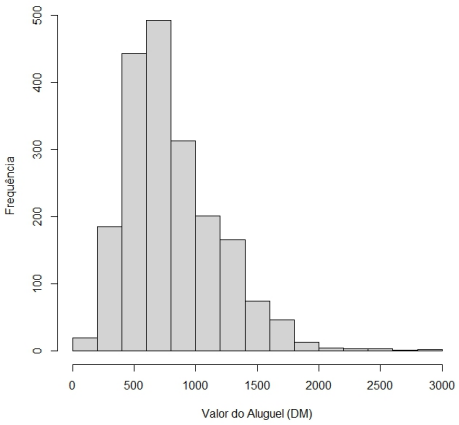
Tabela 3 – Estatísticas descritivas da variável `R`.

	Min	**1ºQ**	**Mediana**	**Média**	**3ºQ**	**Max**	**D.P.**
Aluguel Mensal (em MD)	101,7	544,2	737,8	811,9	1022	3000	379

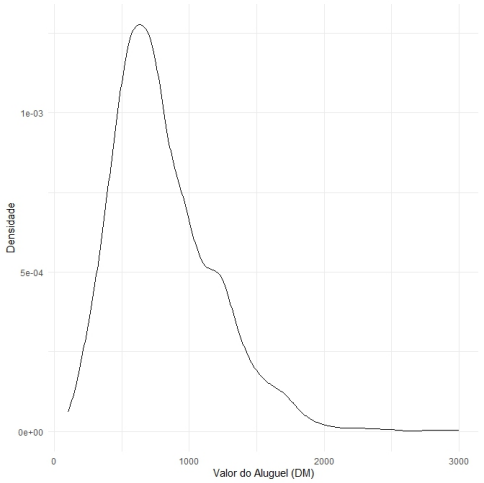
Nas Figuras 2a e 2b, têm-se um histograma e a densidade da variável `R`, respectivamente. Nota-se que, aparentemente, existe uma assimetria à direita, formando

uma espécie de cauda. Já nas Figuras 3a e 3b, vê-se o gráfico de dispersão e um box plot dessa mesma variável, respectivamente. Percebe-se que cerca de 75% das observações estão concentradas entre 0 e 1000 DM, levando em consideração o valor do 3º quartil da Tabela 3. Além disso, é possível identificar a presença de alguns outliers no box plot.

Figura 2 – Análise gráfica da variável `R`.

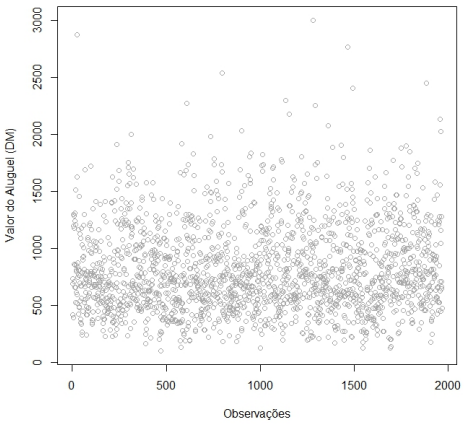


a) Histograma

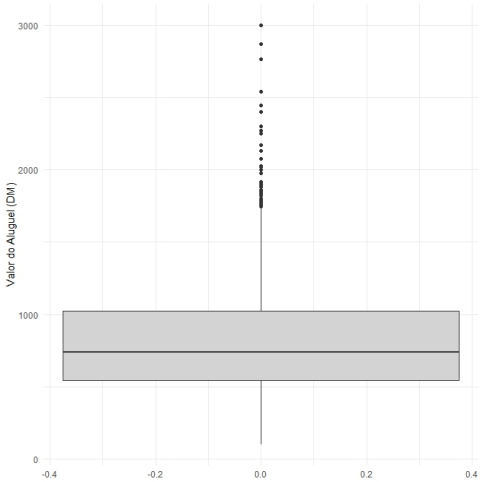


(b) Densidade

Figura 3 – Análise gráfica da variável `R`.



a) Diagrama de dispersão



b) Box plot

A Tabela 4 mostra algumas estatísticas descritivas da variável `Fl`, por ela é correto afirmar que o menor apartamento tem 30 m^2, e o maior tem 120 m^2; os demais estão dentro deste intervalo. Ademais, com a Tabela 5, pode-se afirmar que mais de 70% dos apartamentos observados têm entre 50 m^2 e 101 m^2.

Tabela 4 – Estatísticas descritivas da variável `Fl`.

	Min	**Média**	**Max**
Área (m^2)	30	67	120

Tabela 5 – Contagem de apartamentos por intervalos.

	<= 50 m^2	**>50 e <101** m^2	**>100** m^2
Contagem	452 (23,05%)	1389 (70,83%)	120 (6,12%)

Seguindo com a variável independente `A`, que representa o ano de construção do imóvel, de acordo com os dados observados, o apartamento mais antigo foi constuído no ano de 1890 e os mais novos em 1988. No total, foram registrados 73 anos entre 1890 e 1988. Na Tabela 6, é possível visualizar os 5 anos que tiveram mais registros de apartamentos, em ordem decrescente.

Tabela 6 – Maiores frequências de apartamentos construídos por ano.

Ano de Construção	**Número de Apto**
1893	305
1934	226
1957	551
1972	364
1981	120

Agora, analisando a variável `H`, que indica se o apartamento tem aquecimento central ou não, vê-se, na Tabela 7, que a grande maioria dos apartamentos tem aqueciemnto central.

Tabela 7 – Distribuição dos apartamentos que possuem aquecimento central.

Aquecimento Central	**Frequência**
Sim	1580
Não	389

Para variável `B`, que informa sobre ter ou não banheiro premium no imóvel, verifica-se na Tabela 8, que a grande maioria dos apartamentos possui banheiro (97,77%).

Tabela 8 – Distribuição dos apartamentos que possuem banheiro.

Banheiro Premium	**Frequência**
Sim	1925 (97,77%)
Não	44 (2,23%)

Por fim, para variável `loc`, que mostra a qualidade da localização do apartamento, tem-se a Tabela 9, a qual indica que a grande maioria dos apartamentos, 63,33%, estão categorizados como estando na média

Tabela 9 – Distribuição da qualidade de localização do apartamento.

Qualidade de Localização	**Frequência**
Abaixo da média	172 (8,74%)
Na média	1247 (63,33%)
Acima da média	550 (27,93%)

Com isso, foram feitas analises em conjunto da variável resposta `R` com as 4 variáveis explicativas estudadas.

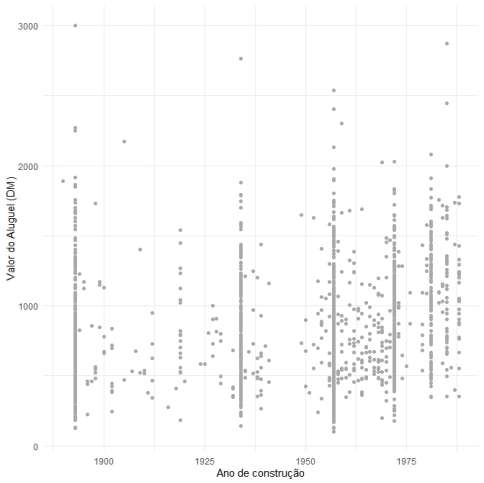
As Figuras 4a e 4b ilustram o comportamento da variável `R` com as variáveis `Fl` e `A`, respectivamente. Observa-se que quanto maior o apartamento, maior o valor do aluguel, com alta variabilidade, fazendo com que exista uma suposta violação de homogeneidade de variância. Já para o ano de construção, percebe-se a complexidade dos dados, nota-se que existe uma tendência de aumento da renda média ao longo dos anos.

Na Figura 5a está apresentado um box plot da variável `H`. Nele, observa-se a variação no valor do aluguel quando as categorias são levadas em consideração. Vê-se que os apartamentos que possuem aquecimento central têm uma mediana superior aos que não possuem, além disso, é possível visualizar que o grupo que possui aquecimento tem mais valores discrepantes que o grupo que não possui. E ainda, percebe-se a existência de assimetria à direita no grupo que possui aquecimento central no apartamento.

A Figura 5b ilustra um box plot da variável `B`. Nota-se que a mediana do grupo que possui banheiro premium é superior ao que não possui. Ademais, esse grupo também apresentou muitos valores discrepantes dentro dos dados.

Figura 4 – Gráficos da variável resposta `R` com as variáveis explicativas numéricas.

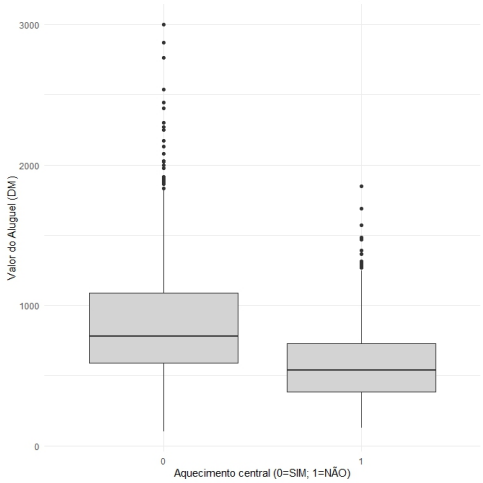
a) Diagrama de dispersão `R` vs `Fl`



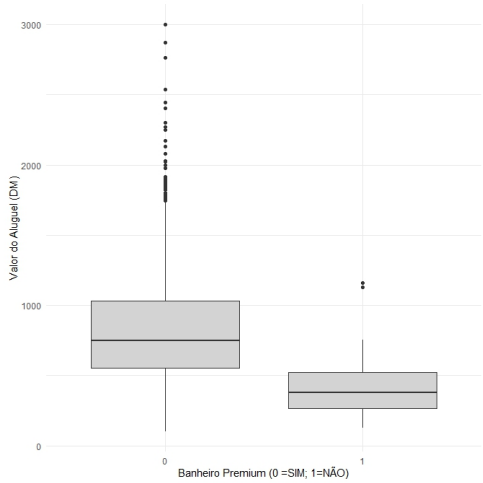
(b) Diagrama de dispersão `R` vs `A`

Já na Figura 6a, é visto o comportamento de 3 fatores relacionados à variável do valor do aluguel. Nota-se que o valor da mediana para a categoria 2 ficou entre as outras duas medianas. Também, vê-se que existe uma assimetria à direita nos grupos, em que a categoria 2 teve mais valores discrepantes. Diante disso, pode-se dizer que a mediana aumenta se o apartamento tiver aquecimento central e aumenta à medida que a localização melhora a qualidade.

Figura 5 – Gráficos da variável resposta `R` com as variáveis explicativas com dois fatores.

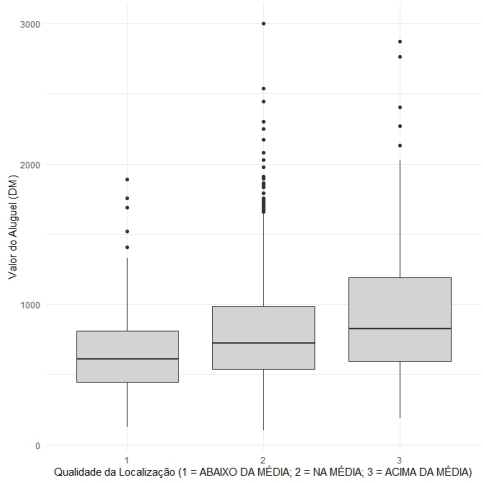


a) Box plot de 2 fatores `R` vs `H`



b) Box plot de 2 fatores `R` vs `B`

Figura 6 – Análise do box plot de 3 fatores `R` vs `loc`.



5.3 Seleção de Variáveis Explicativas da Base `rent`

Nesta seção, foram implementados os métodos de seleção de variáveis explicativas, isto é, a Estratégia A, o critério GAIC e os testes de Vuong e Clarke, até chegar a um modelo final.

5.3.1 Estratégia A

5.3.1.1 Etapa Linear

- i) Ajustar o modelo normal considerando todas as variáveis explicativas para todos os parâmetros da distribuição e usar a função `stepGAICAll.A()` para $k = 2$ e $k = \log(n)$.

Neste tópico, foi ajustado um modelo GAMLSS considerando `Fl`, `A`, `B`, `H` e `loc` como variáveis explicativas e `R` como variável resposta. Em seguida, foi utilizada, duas vezes, a função `stepGAICALL.A()` para o modelo, atribuindo todas variáveis para os escopos de μ e de σ; na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$, em que $n = 1969$.

Observou-se que, após a aplicação da função `stepGAICALL.A()`, para $k = 2$, os parâmetros μ e σ foram modelados com todas as variáveis explicativas, entretanto, para $k = log(n)$, o parâmetro μ foi modelado com todas variáveis explicativas e σ recebeu apenas duas variáveis (`Fl` e `H`).

- ii) Ajustar o modelo BCCGo considerando todas as variaveis explicativas para todos os parâmetros da distribuição e usar a função `stepGAICAll.A()` com $k = 2$ e $k = log(n)$.

Neste tópico, foi ajustado um modelo GAMLSS considerando `Fl`, `A`, `B`, `H` e `loc` como variáveis explicativas e `R` como variável resposta. Em seguida, foi utilizada, duas vezes, a função `stepGAICALL.A()` para o modelo, atribuindo todas variáveis para os escopos de μ, σ e de ν; na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$, em que $n = 1969$.

Quando aplicada a função `stepGAICALL.A()` ao modelo considerando a distribuição BCCGo, o retorno obtido é que para $k = 2$, todos os parâmetros da distribuição foram modelados, sendo μ explicado por todas variáveis; σ sendo explicado por duas variáveis (`A` e `loc`); e ν sendo explicado também por duas variáveis (`A` e `H`). Todavia, quando $k = log(n)$, apenas dois parâmetros foram modelados, sendo μ explicado por todas variáveis e σ sendo explicado por apenas uma variável (`A`).

5.3.1.2 Etapa Com Termo de Suavização

- i) Ajustar o modelo normal considerando todas as variáveis explicativas para todos os parâmetros da distribuição considerar `pb(Fl)` e `pb(A)` e usar a função `stepGAICAll.A()` com $k = 2$ e $k = log(n)$.

Agora, considerando a função de suavização `pb()`, para as variáveis respostas numéricas `Fl` e `A`, foi ajustado um modelo GAMLSS considerando as mesmas variáveis explicativas e a variável resposta da etapa linear. Em seguida, foi utilizada, duas vezes, a função `stepGAICALL.A()` para o modelo, atribuindo todas variáveis para os escopos de μ e de σ, considerando a suavização; na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$, em que $n = 1969$.

Para o modelo normal com suavização com $k = 2$, a função `stepGAICALL.A()` mostrou que um possível modelo que teria um bom ajuste tem μ e σ modelados, tendo todas variáveis modelando a média e quatro modelando a variância (`Fl`, `H`, `A` e `B`); com

$k = log(n)$, os dois parâmetros também foram modelados, porém σ recebeu três variáveis (`Fl`, `H` e `A`).

- i) Ajustar o modelo BCCGo considerando todas as variáveis explicativas para todos os parâmetros da distribuição considerar `pb(Fl)` e `pb(A)` e usar a função `stepGAICAll.A()` com $k = 2$ e $k = log(n)$.

Aqui, também foi aplicada a função de suavização `pb()` nas as variáveis explicativas numéricas `Fl` e `A`; em seguida, foi ajustado um modelo GAMLSS considerando todas variáveis explicativas (com o termo de suavização para as especificadas) e a variável resposta. Por fim, foi utilizada, duas vezes, a função `stepGAICALL.A()` para o modelo, atribuindo todas variáveis para os escopos de μ, σ e de ν, considerando a suavização; na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$, em que $n = 1969$.

Para o modelo considerando a distribuição BCCGo e a suavização `pb()`, para $k = 2$, todos os parâmetros foram modelados, em que μ recebeu todas as variáveis explicativas, σ recebeu quatro variáveis (`A`, `loc` e `Fl`) e ν recebeu duas variáveis (`H` e `A`). Contudo, para $k = log(n)$, apenas dois parâmetros foram modelados em função de variáveis explicativas, sendo μ explicado por todas variáveis e σ sendo explicado apenas pela variável `A`.

5.3.2 Seleção de Modelo `rent`

5.3.2.1 Função `GAIC()`

O primeiro passo para seleção de modelo foi verificar os valores do GAIC para os modelos implementados com cada uma das distribuições. De acordo com a Tabela 10, o modelo com o menor valor do GAIC dentre os modelos ajustados com a distribuição normal é o `M_norm_suav_k2`, isto é, modelo ajustado normal suavizado, considerando o argumento $k = 2$. Por outro lado, em se tratando dos modelos ajustados com a distribuição BCCGo, o que tem o menor valor do GAIC é o `M_bccgo_suav_k2`, ou seja, modelo ajustado com BCCGo suavizado, considerando o argumento $k = 2$.

Tabela 10 – Valores do GAIC para os modelos ajustados com as distribuições normal e BCCGo.

Modelos normal	**df**	**AIC**
M_norm_suav_k2	18,71793	27766,48
M_norm_suav_k_logn	17,71832	27766,95
M_norm_k2	14,00000	27884,84
M_norm_k_logn	10,00000	27890,20
M_norm_suav	12,74285	28079,74
M_norm	8,00000	28169,88

Modelos BCCGo	**df**	**AIC**
M_bccgo_suav_k2	25,71778	27586,09
M_bccgo_suav_k_logn	16,35784	27614,43
M_bccgo_suav	12,40312	27674,19
M_bccgo_k2	14,00000	27710,37
M_bccgo_k_logn	10,00000	27717,35
M_bccgo	9,00000	27752,96

5.3.2.2 Função `VC.test()`

Por fim, para selecionar um entre os dois modelos candidatos, da Seção 5.3.2.1, foi implementada a função `VC.test()`. Portanto, pelos testes de Vuong e Clarke, o modelo `M_bccgo_suav_k2` foi o escolhido. Para esse modelo foi utilizada a função de ligação logarítmica para μ e σ, e a função identidade para o parâmetro ν.

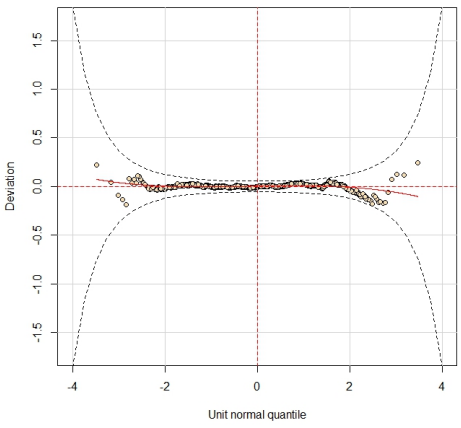
5.3.2.3 Análise de Resíduos

Aqui foram feitas análises gráficas dos resíduos dos modelos selecionados na Seção 5.3.2.1. A Figura 7 apresenta dois gráficos *worm plot*, enquanto a Figura 8 apresenta dois gráficos *bucket plot*, sendo um para cada modelo candidato.

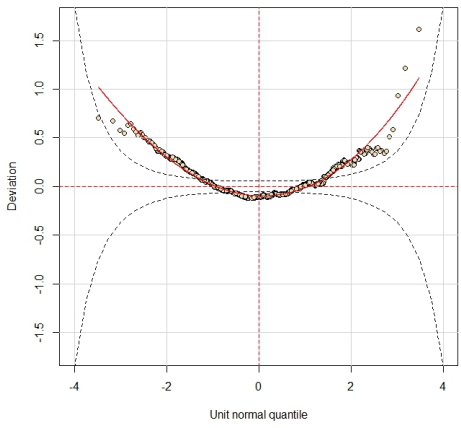
A Figura 7a exibe o *worm plot* do modelo `M_bccgo_suav_k2`, sugerindo uma distribuição normal dos resíduos, conforme indicado pela concentração dos pontos dentro do intervalo de confiança de 95% e ao longo da linha pontilhada. Ademais, o *bucket plot* na Figura 8a reforça essa ideia, pois os pontos estão concentrados em torno das coordenadas $(0, 0)$. Isso sugere que o modelo é satisfatório em relação à assimetria e curtose.

Na Figura 7b, que representa o modelo `M_norm_suav_k2`, observa-se uma forma parabólica com vários pontos fora do intervalo de confiança de 95%, indicando uma má adequação do ajuste. Isso sugere assimetria, curtose e não normalidade nos resíduos. O *buckt plot* correspondente, na Figura 8b, mostra uma concentração de pontos no lado superior direito, evidenciando assimetria à direita e curtose do tipo leptocúrtica.

Figura 7 – Análise gráfica dos resíduos.

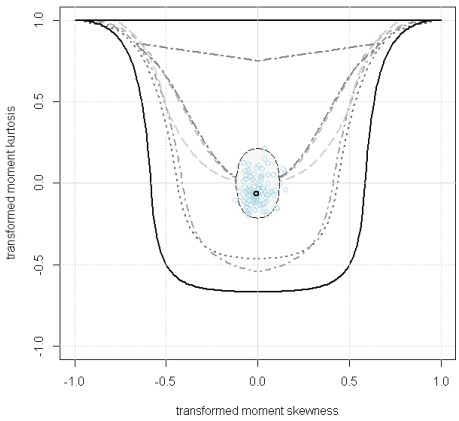


a) *Worm plot* do modelo `M_bccgo_suav_k2`

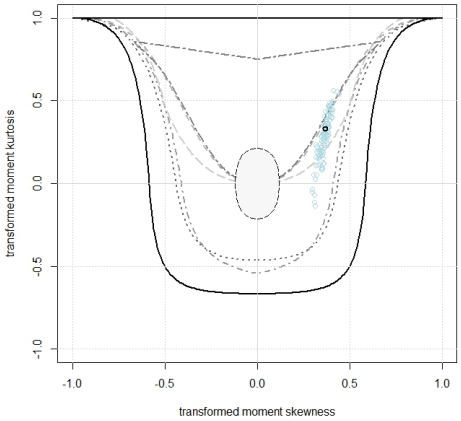


b) *Worm plot* do modelo `M_norm_suav_k2`

Figura 8 – Análise gráfica dos resíduos.



a) *Buckt plot* do modelo `M_bccgo_suav_k2`



b) *Bucket plot* do modelo `M_norm_suav_k2`

5.3.3 Modelo Final `rent`

O diferencial dos testes de Vuong e Clarke é que são apropriados para testar modelos que não necessariamente estejam aninhados. Por isso, o modelo ajustado selecionado por esses testes mostra-se mais adequado e satisfatório para modelar o aluguel mensal da base `rent`. O modelo final é dado por

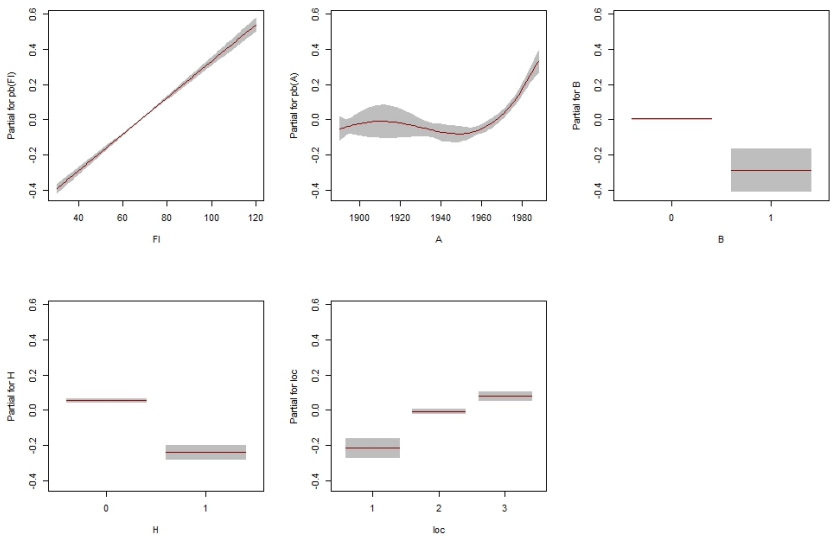
$$\begin{aligned} Y &\sim BCCGo(\mu, \sigma, \nu) \\ \log(\hat{\mu}) &= 2,367317 + 0,010321 \times pb(Fl) + 0,001747 \times pb(A) - 0,292572 \times B_1 \\ &\quad - 0,296243 \times H_1 + 0,208387 \times loc_2 + 0,295586 \times loc_3; \\ \log(\hat{\sigma}) &= 6,709242 - 0,003965 \times pb(A) + 0,001552 \times pb(Fl) - 0,111281 \times loc_2 \\ &\quad - 0,171751 \times loc_3; \\ \hat{\nu} &= -3,425685 - 0,252241 \times H_1 + 0,002012 \times pb(A); \end{aligned} \tag{5.1}$$

em que `pb` é uma função de suavização não paramétrica.

Para analisar a contribuição aditiva dos termos na modelagem dos parâmetros, foi utilizada a função `term.plot()`. As Figuras 9–11 apresentam os termos paramétricos ajustados no modelo final escolhido (5.1). Nelas são mostrados os termos em $\log(\hat{\mu})$, $\log(\hat{\sigma})$ e $\hat{\nu}$.

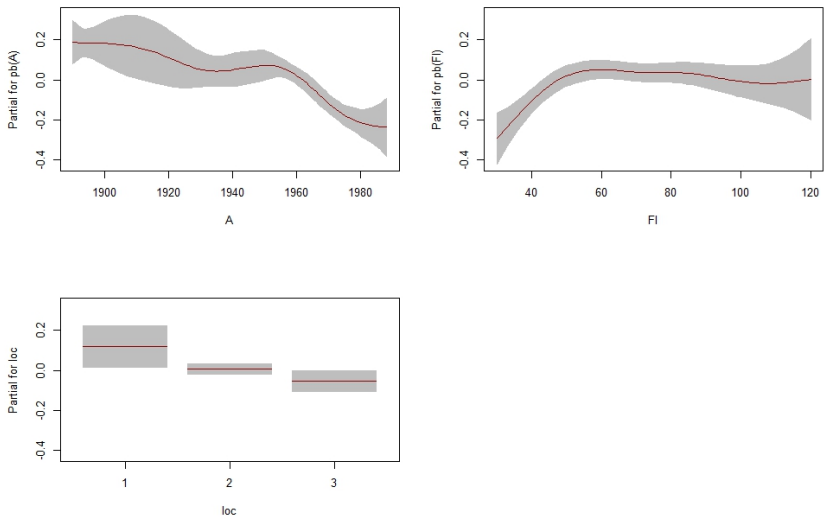
A Figura 9 expõe os termos paramétricos ajustados em $\log(\hat{\mu})$. O efeito da mediana ajustada $\hat{\mu}$ do aluguel mensal dos apartamentos parcial aumenta linearmente com a área do imóvel (`Fl`) e, de forma não linear, com o ano de construção (`A`). Em contrapartida, o efeito da mediana diminui quando o apartamento não tem banheiro premium (`B`), não possi aquecimento central (`H`) e está numa localização abaixo da média (`loc`).

Figura 9 – *Term plots* para $\log(\hat{\mu})$.



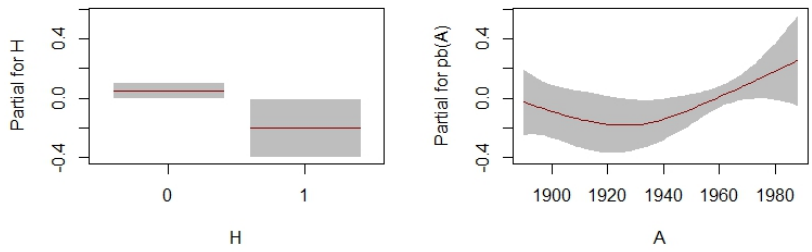
A Figura 10 mostra os termos paramétricos ajustados em $\log(\hat{\sigma})$. O efeito de $\hat{\sigma}$ tem um decaimento não linear com o ano de construção (`A`) e um aumento com a área do apartamento (`Fl`), se mantendo constante para áreas acima de 50 metros quadrados. Ademais, o efeito tem uma diminuição com a localização de apartamentos acima da média (`loc`).

Figura 10 – *Term plots* para $\log(\hat{\sigma})$.



Na Figura 19,está exibido o termo paramétrico ajustado em ν, mostrando que o efeito de $\hat{\nu}$ aumenta com o ano de contrução (`A`), mas diminui quando o apartamento não possui aquecimento central (`H`).

Figura 11 – *Term plots* para $\hat{\nu}$.



5.4 Análise Descritiva da Base `wines`

Nesta seção, foi realizada uma análise exploratória das variáveis utilizadas para modelar o preço da garrafa de vinho do conjunto de dados `wines`. A descrição desses dados pode ser vista na Seção 5.1.2. Aqui, busca-se entender o comportamento dos dados e obter informações relevantes para o ajuste do modelo.

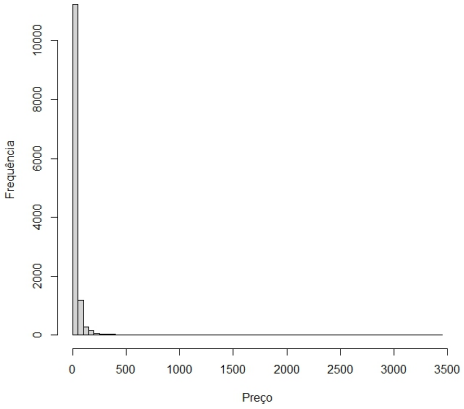
A Tabela 11 mostra algumas estatísticas da variável resposta `Price`. Pode-se observar que há uma grande diferença entre o maior e o menor preço, em que o vinho mais barato custa €3,55 e o mais caro custa €3.410,79.

Tabela 11 – Estatísticas do preço dos vinhos, em Euros.

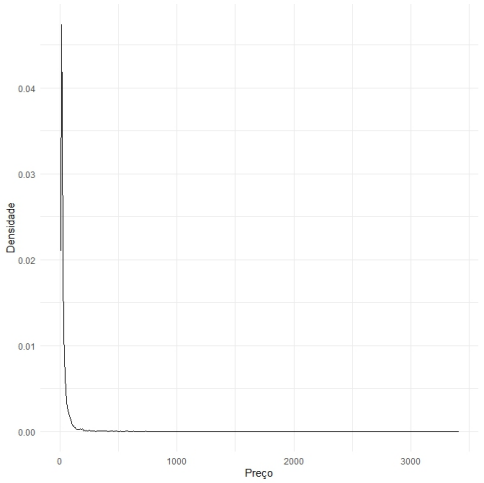
	Mínimo	**1º Q**	**Mediana**	**Média**	**3º Q**	**Máximo**
Preço	3,55	9,9	15,95	33,49	32,5	3410,79

As Figuras 12a e 12b exibem, respectivamente, um histograma e a densidade da distribuição de preços de vinhos revelando uma assimetria acentuada à direita. Já as Figuras 12c e 12d mostram um diagrama de dispersão e um box plot dos preços, evidenciando uma grande variabilidade nos dados, além de diversos valores discrepantes. Destaca-se um valor significativamente distante dos demais, ultrapassando €3.000,00, enquanto a maioria dos outros valores permanece abaixo de €2.000,00.

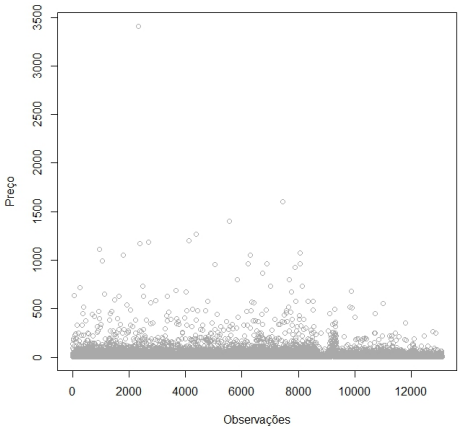
Figura 12 – Análise gráfica da variável `Price`.



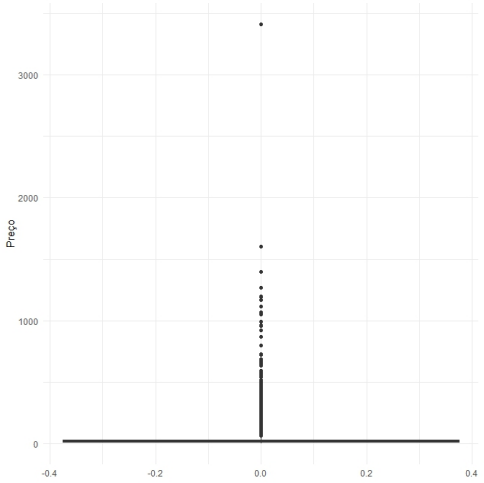
a) Histograma



(b) Densidade



c) Diagrama de dispersão



d) Box plot

Seguindo para a variável `Year`, observa-se que a grande maioria dos vinhos analisados foi produzida durante o período compreendido entre os anos 2000 e 2020. Além disso, é interessante notar que os três vinhos mais antigos na base de dados datam o ano de 1961, o que pode ser relevante para análises históricas ou de envelhecimento dos vinhos.

Ao considerar a classificação dos vinhos em quatro tipos distintos, conforme ilustrado na Tabela 12, destaca-se que a maioria das observações, aproximadamente 66%, corresponde ao vinho Tinto. O segundo tipo mais frequente é o vinho Branco, representando cerca de 28,7% das observações. Os tipos Rosé e Espumante descrevem uma proporção menor, com aproximadamente 3% e 2,1% das observações, respectivamente.

Tabela 12 – Tipos de vinhos.

Tipo	Branco	Espumante	Rosé	Tinto
Frequência	3759 (28,72%)	279 (2,13%)	394 (3,01%)	8658 (66,14%)

Na Tabela 13, está apresentada a frequência de cada país de origem. Pode-se

observar que os países europeus Itália, França, Espanha e Alemanha têm as maiores frequências na base de dados estudada. Também, é possível notar que o 49 vinhos foram fabricados no Brasil.

Tabela 13 – Frequência dos países de origem.

País	Frequência	País	Frequência	País	Frequência
Itália	3636	Nova Zelândia	164	Turquia	10
França	3144	Brasil	49	Croácia	6
Espanha	1483	Romênia	37	Luxemburgo	5
Alemanha	1176	Grécia	23	Uruguai	5
África do Sul	833	Israel	23	Canadá	3
Estados Unidos	525	Suíça	23	China	3
Áustria	477	Hungria	19	Reino Unido	3
Chile	429	Eslovênia	17	Bulgária	2
Portugal	327	Líbano	16	República Checa	2
Austrália	313	Geórgia	13	Eslováquia	2
Argentina	308	Moldávia	13	México	1

A avaliação média (`Rating`) varia de 1 até 5, sendo 1 a avaliação mínima e 5 a avaliação máxima. Essa nota é calculada a partir da média aritmética das avaliações feitas pelos consumidores, referentes à satisfação pelo produto, isto é, cada cliente deixou uma avaliação no site, classificando o produto com as notas 1, 2, 3, 4 ou 5, e por meio dessas notas, foi calculada a média aritmética da variável, que inclusive já estava disponível na base de dados. A Tabela 14 expõe algumas estatísticas sobre a avaliação média dos vinhos. Verifica-se que a menor avaliação média foi de 2,5 e a maior foi de 4,9, nenhum vinho teve a avaliação média máxima ou mínima (1 e 5, respectivamente). A média das avaliações médias foi de, aproximadamente, 3,87.

Tabela 14 – Estatísticas da avaliação média dos vinhos.

	Mínimo	**1º Q**	**Mediana**	**Média**	**3º Q**	**Máximo**
Avaliação	2,5	3,7	3,9	3,869	4,1	4,9

Na Tabela 15, vê-se que o vinho mais caro da base de dados estudada foi do tipo Tinto produzido na França, em 2012, que possui uma avaliação média de 4,7 e um valor de €3.410,79; vale ressaltar que ele foi avaliado por 204 consumidores. Por outro lado, o vinho mais barato também foi um vinho tinto, entretanto ele foi fabricado na Espanha, em 2018, com avaliação média igual a 3,2, custando €3,55 e tendo sido avaliado por 44 pessoas.

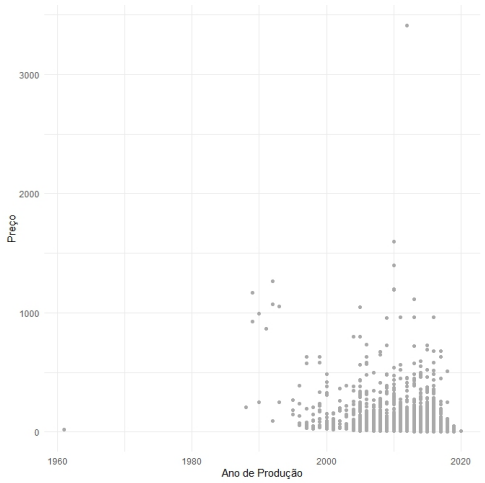
Tabela 15 – Estatísticas dos vinhos.

	Vinho	**Preço**	**País**	**Ano**	**Av. Média**	**Número de Avaliações**	**Nome do Vinho**
Mais caro	Red	3410,79	França	2012	4,7	204	Pomerol
Mais barato	Red	3,55	Espanha	2018	3,2	44	Shiraz

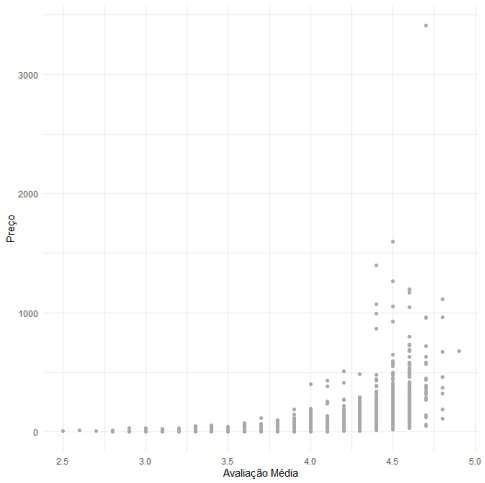
A Figura 13a ilustra o comportamento da relação entre a variável preço e ano de produção. Repara-se que os dados são muito disperso, existindo grande variabilidade.

Por outro lado, a Figura 13b mostra a relação da variável preço com a avaliação média; constata-se que os vinhos mais avaliados tendem a ser mais caros.

Figura 13 – Análise gráfica da variável resposta `Price` com as variáveis explicativas numéricas.



a) Diagrama de dispersão `Price` vs `Year`

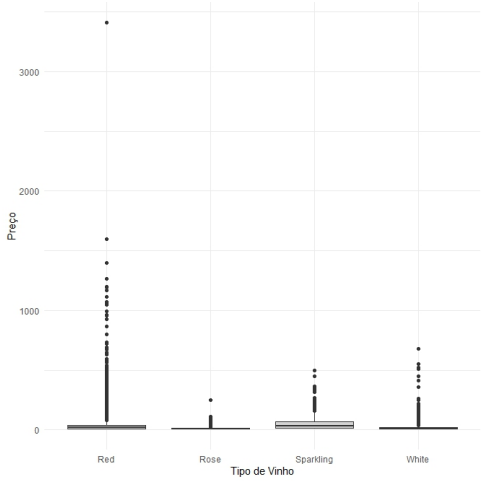


(b) Diagrama de dispersão `Price` vs `Rating`

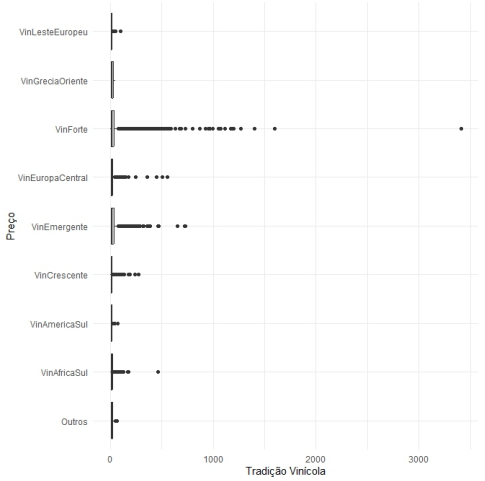
A Figura 14a ilustra a relação entre os tipos de vinhos e os preços. Observa-se que a mediana do tipo Tinto e Espumante parecem maior. Além disso, é nítida a presença de muitos valores discrepantes para o vinho Tinto. Mais ainda, percebe-se uma assimetria no grupo dos vinhos do tipo Espumante.

A Figura 14b apresenta um box plot da relação entre as tradições vinícolas e o preço dos vinhos. A figura evidencia que as tradições vinícolas Forte, Emergente e Grécia Oriente parecem ter distribuições assimétricas. Mais ainda, vale destacar a presença de diversos outliers na tradição vinícola Vinho Forte.

Figura 14 – Análise gráfica da variável resposta `Price` com as variáveis explicativas.



a) Box plot de 4 fatores `Price` vs `Ft`



b) Box plot de 9 fatores `Price` vs `Fc`

5.5 Seleção de Variáveis Explicativas da Base `wines`

A análise exploratória na seção anterior foi importante para a compreender melhor a distribuição da variável referente ao preço dos vinhos. Dada a complexidade dos dados, agora o avanço se dá para a etapa da seleção do modelo estatístico que melhor se ajusta aos dados estudados. O método de seleção para este conjunto de dados é análogo ao método utilizado na Seção 5.3.2.

5.5.1 Estratégia A

5.5.1.1 Etapa Linear

- i) Ajustar o modelo BCPEo considerando todas as variáveis explicativas para todos os parâmetros da distribuição e usar a função `stepGAICAll.A()` para $k = 2$ e $k = \log(n)$.

Neste tópico, foi ajustado um modelo GAMLSS considerando `Rating`, `Year`, `Fc`, `Ft` como variáveis explicativas e `Price` como variável resposta. Em seguida, foi utilizada, duas vezes, a função `stepGAICALL.A()` para o modelo, atribuindo todas variáveis para os escopos de μ, σ, ν e τ; na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$, em que $n = 13090$.

Após a aplicação da função `stepGAICALL.A()`, para $k = 2$, o parâmetro μ foi modelado com todas as variáveis explicativas e τ foi modelado com apenas uma variável (`Rating`). O mesmo aconteceu para $k = log(n)$, o parâmetro μ foi modelado com todas variáveis explicativas e τ foi modelado com a variável `Rating`.

- ii) Ajustar o modelo BCTo, considerando todas as variaveis explicativas para todos os parâmetros da distribuição e usar a função `stepGAICAll.A()` com $k = 2$ e $k = log(n)$.

Aqui também foi ajustado um modelo GAMLSS considerando `Rating`, `Year`, `Fc`, `Ft` como variáveis explicativas e `Price` como variável resposta. Em seguida, foi implementada, duas vezes, a função `stepGAICALL.A()` para o modelo, atribuindo todas variáveis para os escopos de μ, σ, ν e τ; na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$. No retorno obtido, para $k = 2$, apenas os parâmetros μ e τ foram modelados, em que μ foi explicado por todas variáveis e τ foi explicado por 3 variáveis (`Rating`, `Year` e `Ft`). Todavia, para $k = \log n$, o resultado foi semelhante, contudo τ foi modelado apenas com as variáveis `Rating` e `Year`.

5.5.1.2 Etapa Com Termo de Suavização

- i) Ajustar o modelo BCPEo considerando todas as variáveis explicativas para todos os parâmetros da distribuição considerar `pb(Rating)` e `pb(Year)` e usar a função

`stepGAICAll.A()` com $k = 2$ e $k = log(n)$.

Considerando a função de suavização `pb()`, para as variáveis respostas numéricas `Rating` e `Year`, foi ajustado um modelo GAMLSS considerando todas as variáveis explicativas. Posteriormente, foi implementada duas vezes a função `stepGAICALL.A()` para o modelo, atribuindo todas variáveis para os escopos de μ, σ, ν e τ, levando em consideração o termo de suavização. Na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$.

Após a implementação, a função `stepGAICALL.A()`, para o argumento $k = 2$, retornou que todos os parâmetros foram modelados, em que μ recebeu todas as variáveis explicativas; σ recebeu duas variáveis (`Year` e `Fc`); ν recebeu três variáveis (`Rating`, `Year` e `Ft`); e o argumento do parâmetro ν recebeu apenas a variável `Fc`. Por outro lado, para $k = \log n$, apenas μ foi modelado com todas as variáveis explicativas.

- i) Ajustar o modelo BCTo considerando todas as variáveis explicativas para todos os parâmetros da distribuição, considerando `pb(Rating)` e `pb(Year)` e usar a função `stepGAICAll.A()` com $k = 2$ e $k = log(n)$.

Analogamente, foi aplicada a função de suavização `pb()` nas as variáveis respostas numéricas `Rating` e `Year`; em seguida, foi ajustado um modelo GAMLSS, considerando todas variáveis explicativas; por fim, foi utilizada, duas vezes, a função `stepGAICALL.A()` para o modelo, atribuindo todas as variáveis para os escopos de μ, σ, ν e τ, considerando a suavização; na primeira busca foi considerado o valor de $k = 2$ e, na segunda, o valor de $k = log(n)$.

Para $k = 2$, retornou que todos os parâmetros foram modelados, em que μ e σ receberam todas as variáveis explicativas; o parâmetro ν recebeu duas variáveis (`Rating` e `Year`); e τ recebeu três variáveis (`Rating`, `Year` e `Fc`). Porém, para $k = \log n$, apenas μ, σ e ν foram modelados, em que os argumentos μ e σ receberam todas as variáveis explicativas e ν recebeu apenas a variável `Rating`.

5.5.2 Seleção de Modelo `wines`

5.5.2.1 Função `GAIC()`

Análogo à Seção 5.3.2, o primeiro passo para escolher o modelo mais adequado foi observar os valores do GAIC para os modelos implementados com cada uma das distribuições. Conforme a Tabela 16, o modelo com o menor valor do GAIC dentre os modelos ajustados com a distribuição BCPEo é o `M_bcpeo_suav_k2`, isto é, modelo ajustado BCPEo suavizado, considerando o argumento $k = 2$. Por outro lado, dentre os modelos ajustados com a distribuição BCTo, o que tem o menor valor do GAIC é o `M_bcto_suav_k2`, ou seja, modelo ajustado com BCTo suavizado, considerando o argumento $k = 2$.

Tabela 16 – Valores do GAIC para os modelos ajustados com as distribuições BCPEo e BCTo.

Modelos BCPEo	**df**	**AIC**
M_bcpeo_suav_k2	72,03641	93645,89
M_bcpeo_suav	33,36474	94228,11
M_bcpeo_suav_k_logn	33,36474	94228,11
M_bcpeo_k2	18,00000	96774,95
M_bcpeo_k_logn	18,00000	96774,95
M_bcpeo	17,00000	96796,16

Modelos BCTo	**df**	**AIC**
M_BCTo_suav_k2	72,89030	92983,41
M_BCTo_suav_k_logn	56,73515	93081,65
M_BCTo_suav	33,69407	94217,16
M_BCTo_k2	22,00000	96000,90
M_BCTo_k_logn	19,00000	96014,26
M_BCTo	17,00000	96552,94

5.5.2.2 Função `VC.test()`

O último passo para selecionar um entre os dois modelos candidatos, da Seção 5.5.2.1, foi implementar a função `VC.test()`. Portanto, pelos testes de Vuong e Clarke, o modelo `M_bcto_suav_k2` foi preferível em relação ao modelo `M_bcpeo_suav_k2`. Para o modelo escolhido, foi utilizada a função de ligação logarítmica para μ, σ e τ, e a função identidade para o parâmetro ν.

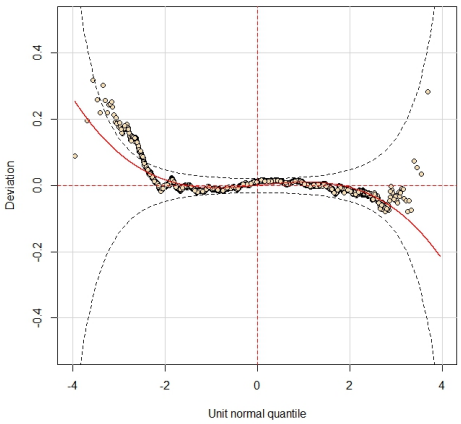
5.5.2.3 Análise de Resíduos

Nesta seção, foi feita uma análise de diagnóstico dos resíduos dos dois modelos candidatos a fim de comparar a adequação do ajuste deles. Na Figura 15, estão ilustrados dois gráficos *worm plot* e, na Figura 16, estão apresentados dois gráficos *bucket plot* de cada modelo candidato.

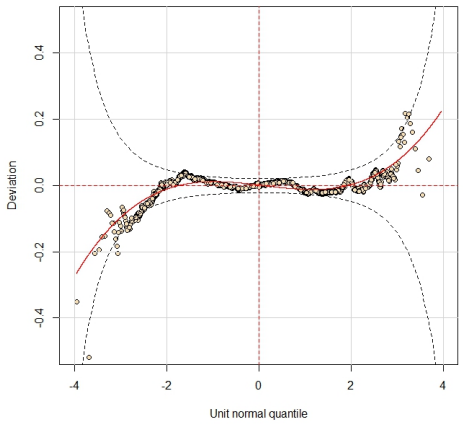
A Figura 15a apresenta um *worm plot* para o modelo `M_bcto_suav_k2`. Embora os resíduos pareçam normalmente distribuídos, há sinais de curtose devido à curvatura da linha vermelha. Além disso, alguns pontos estão fora do intervalo de 95% confiança. Já o *bucket plot*, na Figura 16a, sugere que o modelo é aceitável em relação à assimetria e curtose, com muitos pontos dentro da área circulada e próximos às coordenadas $(0, 0)$.

Na Figura 15b, que mostra o *worm plot* para o modelo `M_bcpeo_suav_k2`, os resíduos parecem distribuir-se normalmente, todavia a curvatura da linha vermelha sugere presença de curtose. Alguns pontos encontram-se fora do intervalo de 95% de confiança. O *bucket plot* subsequente na Figura 16b mostra que a maioria dos pontos está fora da região circulada, indicando uma curtose leptocúrtica.

Figura 15 – Análise gráfica de resíduos.

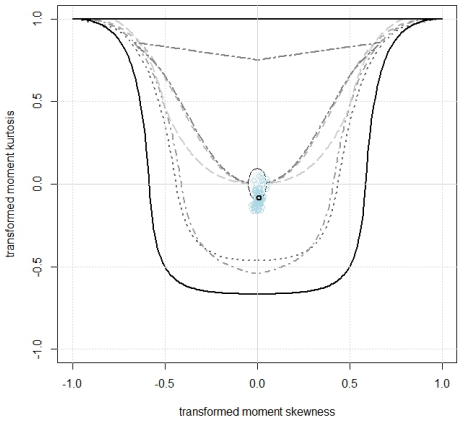


a) *Worm Plot* do modelo `M_bcto_suav_k2`

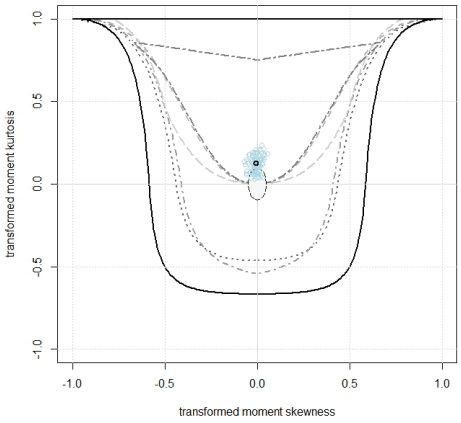


b) *Worm Plot* do modelo `M_bcpeo_suav_k2`

Figura 16 – Análise gráfica de resíduos.



a) *Bucket Plot* do modelo `M_bcto_suav_k2`



b) *Bucket Plot* do modelo `M_bcpeo_suav_k2`

5.5.2.4 Modelo Final `wines`

Com isso, tem-se o necessário para afirmar que o modelo `M_bcto_suav_k2`, selecionado pelos testes de Vuong e Clarke, teve um desempenho melhor. O modelo final é dado por

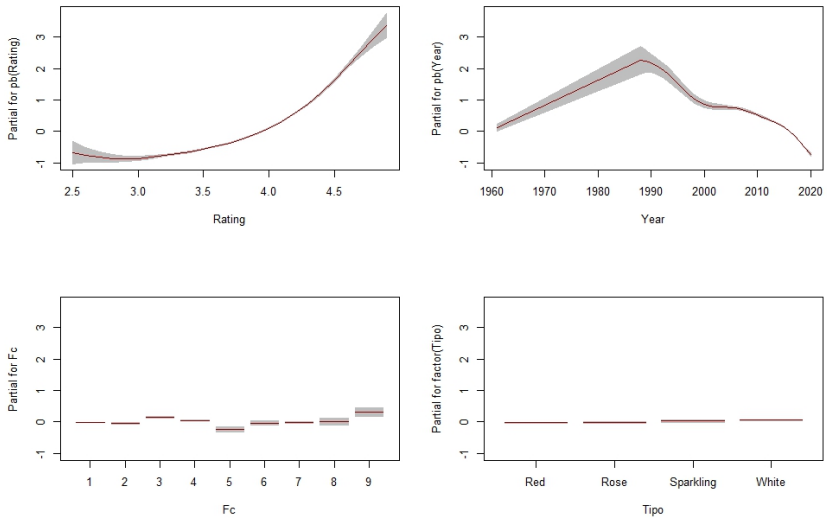
$$
\begin{aligned}
Y &\sim BCTo(\mu, \sigma, \nu, \tau) \\
\log(\hat{\mu}) &= 195,40003 + 1,77023 \times pb(Rating) - 0,09891 \times pb(Year) \\
&\quad - 0,01840 \times Fc_2 + 0,17193 \times Fc_3 + 0,06528 \times Fc_4 \\
&\quad - 0,20223 \times Fc_5 - 0,01020 \times Fc_6 + 0,01105 \times Fc_7 \\
&\quad + 0,03283 \times Fc_8 + 0,34025 \times Fc_9 + 0,01134 \times Ft_2 \\
&\quad + 0,05808 \times Ft_3 + 0,08242 \times Ft_4, \\
\log(\hat{\sigma}) &= -11,139564 + 0,629918 \times pb(Rating) + 0,003936 \times pb(Year) \\
&\quad - 0,355019 \times Fc_2 - 0,079326 \times Fc_3 - 0,179292 \times Fc_4 \\
&\quad - 0,309458 \times Fc_5 - 0,548985 \times Fc_6 - 0,208306 \times Fc_7 \\
&\quad - 0,103270 \times Fc_8 - 0,040009 \times Fc_9 - 0,150615 \times Ft_2 \\
&\quad - 0,277909 \times Ft_3 - 0,084432 \times Ft_4, \\
\hat{\nu} &= 35,15169 + 0,37411 \times pb(Rating) - 0,01832 \times pb(Year) \\
\log(\hat{\tau}) &= -526,3999 - 2,5613 \times Ft_2 + 1,1982 \times Ft_3 - 1,2563 \times Ft_4 \\
&\quad + 3,0760 \times pb(Rating) + 0,2575 \times pb(Year),
\end{aligned}
\tag{5.2}
$$

em que `pb` é uma função de suavização não paramétrica.

Para analisar a contribuição aditiva dos termos na modelagem dos parâmetros, foi utilizada a função `term.plot()`. As Figuras 17–20 apresentam os termos paramétricos ajustados no modelo final escolhido (5.2). Elas exibem os termos em $\log(\hat{\mu})$, $\log(\hat{\sigma})$, $\hat{\nu}$ e $\log(\hat{\tau})$.

A Figura 17 mostra os termos paramétricos ajustados em $\log(\hat{\mu})$. O efeito de $\hat{\mu}$, a mediana ajustada do preço dos vinhos, aumenta com a avaliação média (`Rating`), de forma não linear. No entando a mediana ajustada do preço da garrafa diminui com o passar dos anos (`Year`), indicando que vinhos produzidos nas últimas décadas tendem a ter um preço menor. Além disso, vinhos produzidos nos países (`Fc`) da categoria da tradição vinícola "Outros" resulta em um pequeno aumento no preço, em contrapartida, vinhos produzidos por países da categoria da tradição vinícola "Vinho Leste Europeu" resulta em uma pequena diminuição do preço (as categorias podem ser vistas na Seção 5.1.2.1). Não houve diferenças significativas no efeito da mediana quanto ao tipo do vinho (`Ft`).

Figura 17 – *Term plots* para $\log(\hat{\mu})$.



A Figura 18 mostra os termos paramétricos ajustados em $\log(\hat{\sigma})$. O efeito de $\hat{\sigma}$ tem um aumento não linear com a avaliação média (`Rating`) e com o ano de produção (`Year`), este se mantendo constante depois do ano 1990. Além disso, $\hat{\sigma}$ diminui com vinhos produzidos nos países (`Fc`) da categoria da tradição vinícola "Vinho Forte" mas diminui com os vinhos produzidos nos países da tradição vinícola "Vinho da América do Sul", também diminui quando o vinho é do tipo (`Ft`) espumante.

Na Figura 19,está exibido o termo paramétrico ajustado em ν, mostrando que o efeito de $\hat{\nu}$ se mantém linearmente constante com a avaliação média (`Rating`) e também com o ano de produção (`Year`). Já na Figura 20, que mostra os termos em $\log(\hat{\tau})$, tem-se que o efeito de $\hat{\tau}$ aumenta com a avaliação média, mas não tem muita variação quanto ao tipo de vinhos (`Ft`) e ano de produção.

Figura 18 – *Term plots* para $\log(\hat{\sigma})$.

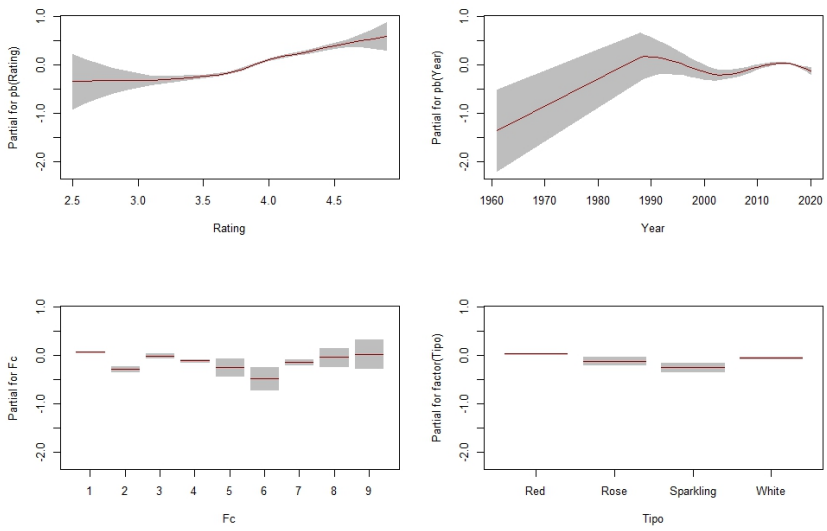


Figura 19 – *Term plots* para $\hat{\nu}$.

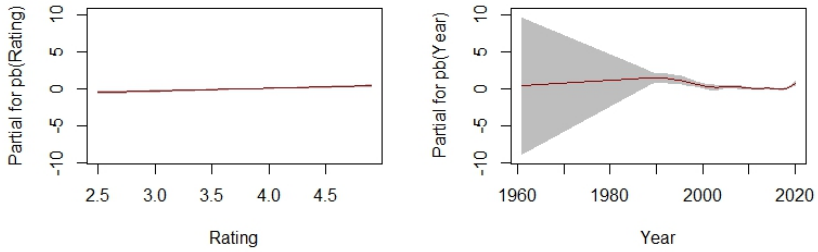
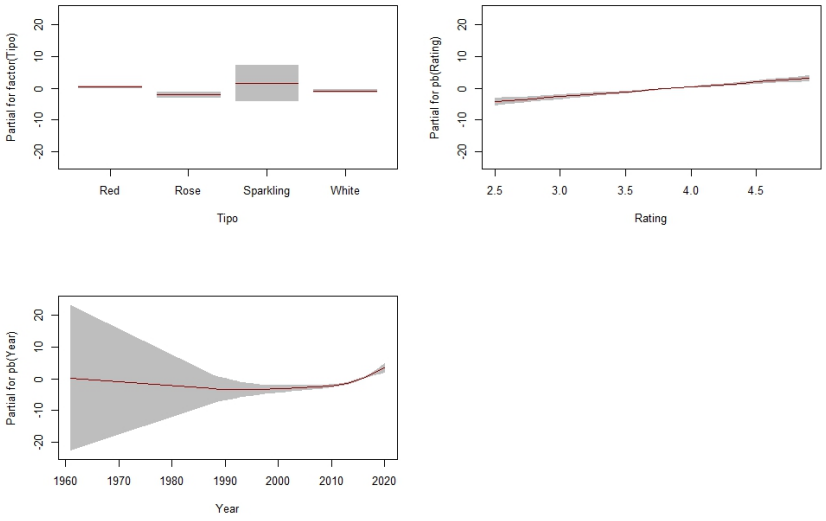


Figura 20 – *Term plots* para $\log(\hat{\tau})$.



6 Conclusão

Neste estudo, foram exploradas e comparadas duas técnicas de seleção de variáveis explicativas nos Modelos Aditivos Generalizados de Locação, Escala e Forma. Foram consideradas diferentes distribuições, duas bases de dados e dois critérios de seleção: GAIC, baseada no AIC; e os testes de Vuong e Clarke, baseados na razão de verossimilhança.

Para construção dos modelos, foi utilizada a Estratégia A, implementada pela função `stepGAICALL.A()` do R. A partir disso, a fim de obter um modelo adequado aos dados dos dois conjuntos de dados estudados, doze modelos ajustados foram testados para cada base, com o argumento $k = 2$ e $k = \log n$. Na base de dados `rent`, foram implementados 6 modelos com a família de distribuição normal e outros 6 com a família de distribuição BCCGo; além disso, foi aplicada a função de suavização `pb()` nas variáveis `A` e `Fl`, em 3 modelos da normal e em 3 modelos da BCCGo. Já para o conjunto de dados `wines`, foram implementados 6 modelos com a família de distribuição BCPEo e outros 6 com a família de distribuição BCTo; ademais, foi aplicada a função de suavização `pb()` nas variáveis `Rating` e `Year`, em 3 modelos da BCPEo e em 3 modelos da BCTo.

Para o conjunto `rent`, pelo critério GAIC, o modelo normal mais adequado foi o `M_norm_suav_k2`, que representa o modelo ajustado suavizado, com o argumento $k = 2$, implementado com a distribuição normal; ademais, o modelo BCCGo mais adequado foi o `M_bccgo_suav_k2`, o qual expressa o modelo ajustado suavizado, com o argumento $k = 2$, implementado com a distribuição BCCGo. Dentre esses dois modelos candidatos, os testes de Vuong e Clarke selecionaram o `M_bccgo_suav_k2` como o modelo que melhor descreve o valor do aluguel mensal. E para verificar a adequação do ajuste, foram utilizados os gráficos de visualização de resíduos *Worm Plot* e *Bucket Plot*, os quais apresentaram resíduos satisfatórios. Para o modelo final (5.1), foram gerados gráficos *term plot* de cada parâmetro estimado para entender a contribuição de cada termo para o modelo.

Para a base de dados `wines`, pelo critério GAIC, o modelo candidato dentre os analisados com a distribuição BCPEo foi o `M_bcpeo_suav_k2`, que designa o modelo ajustado suavizado, considerando o argumento $k = 2$, implementado com a família de distribuição BCPEo. Além disso, o modelo candidato selecionado dentre os avaliados com a distribuição BCTo foi o `M_bcto_suav_k2`, que expressa o modelo ajustado suavizado, considerando o argumento $k = 2$, implementado com a família de distribuição BCTo. Pelos testes de Vuong e Clarke, o modelo preferível foi o `M_bcto_suav_k2`. Os resíduos analisados por meio do *bucket plot* e *worm plot* mostraram-se satisfatórios. Para o modelo final (5.2), foram gerados gráficos *term plot* para entender o feito de cada termo nos parâmetros.

Em síntese, este estudo foi direcionado a fornecer um método mais apropriado para guiar a seleção de variáveis em um contexto dos GAMLSS para modelos não aninhados. Os resultados revelaram que a aplicação da função `VC.test()` proporcionou um modelo

que demonstrou ser adequado às características dos dados de cada uma das bases estudadas. Esse critério de seleção se destacou pela sua eficácia na abordagem de avaliação de modelos não aninhados. A escolha final do modelo foi baseada nos testes de Vuong e Clarke que é respaldada pela sua capacidade de ajustar-se efetivamente à complexidade dos dados.

Para estudos subsequentes a este estudo, proponho que sejam realizadas comparações com outras técnicas de seleção de variáveis explicativas dentro dos GAMLSS. Isso permitirá uma análise mais detalhista e crítica dos resultados obtidos, possibilitando a avaliação da eficácia do método proposto em relação a alternativas já estabelecidas e mais utilizadas, como GAIC, AIC e BIC. A comparação com outras técnicas de seleção de variáveis também proporcionará percepções mais amplas sobre a generalização e aplicabilidade do método em diferentes contextos e conjuntos de dados. Além disso, sugiro que outras funções de suavização sejam implementadas, pois é possível que outro termo de suavização consiga captar com mais precisão a não linearidade entre os dados. Ademais, considerar procedimentos concorrentes aprofundará a compreensão das vantagens e limitações do método proposto, oferecendo uma base para avanços na seleção de variáveis nos GAMLSS.

Referências

AKAIKE, H. Information theory and an extension of the likelihood ratio principle. In: PETROV, B. N.; CSAKI, F. (Ed.). *Second international symposium of information theory*. Budapest: Akademiai Kiado, 1973. (Minnesota Studies in the Philosophy of Science). Citado 2 vezes nas páginas 4 e 12.

AKAIKE, H. Information measures and model selection. *Int Stat Inst*, v. 44, p. 277–291, 1983. Citado na página 4.

BASTIANI, F. D. et al. Gaussian markov random field spatial models in gamlss. *Journal of Applied Statistics*, Taylor & Francis, v. 45, n. 1, p. 168–186, 2018. Citado na página 6.

BASTIANI, F. D. et al. Bucket plot: A visual tool for skewness and kurtosis comparisons. *Brazilian Journal of Probability and Statistics*, Brazilian Statistical Association, v. 36, n. 3, p. 421–440, 2022. Citado 3 vezes nas páginas 2, 6 e 16.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, Taylor & Francis, v. 88, n. 421, p. 9–25, 1993. Citado na página 3.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001. Citado 3 vezes nas páginas 2, 6 e 14.

CLARKE, K. A. A simple distribution-free test for nonnested model selection. *Political Analysis*, Cambridge University Press, v. 15, n. 3, p. 347–363, 2007. Citado 3 vezes nas páginas 2, 6 e 13.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992. Citado na página 4.

DIXIT, S.; JAYAKUMAR, K. A non-stationary and probabilistic approach for drought characterization using trivariate and pairwise copula construction (pcc) model. *Water Resources Management*, Springer, v. 36, n. 4, p. 1217–1236, 2022. Citado na página 6.

HASTIE, T. J. Generalized additive models. In: *Statistical models in S*. [S.l.]: Routledge, 2017. p. 249–307. Citado na página 3.

HOSSAIN, A. et al. Centile estimation for a proportion response variable. *Statistics in medicine*, Wiley Online Library, v. 35, n. 6, p. 895–904, 2016. Citado na página 6.

JARQUE, C. M.; BERA, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, Elsevier, v. 6, n. 3, p. 255–259, 1980. Citado na página 17.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The annals of mathematical statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951. Citado 2 vezes nas páginas 2 e 11.

LIN, X.; ZHANG, D. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 61, n. 2, p. 381–400, 1999. Citado na página 3.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, Oxford University Press, v. 135, n. 3, p. 370–384, 1972. Citado na página 3.

QU, C. et al. Non-stationary flood frequency analysis using cubic b-spline-based gamlss model. *Water*, MDPI, v. 12, n. 7, p. 1867, 2020. Citado na página 5.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>. Citado na página 1.

RAMIRES, T. G. et al. Validation of stepwise-based procedure in gamlss. *Journal of Data Science*, , v. 19, n. 1, p. 96–110, 2021. Citado na página 4.

RIGBY, R. A.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, Springer, v. 6, p. 57–65, 1996. Citado na página 4.

RIGBY, R. A.; STASINOPOULOS, D. M. Smooth centile curves for skew and kurtotic data modelled using the box–cox power exponential distribution. *Statistics in medicine*, Wiley Online Library, v. 23, n. 19, p. 3053–3076, 2004. Citado na página 5.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, Oxford University Press, v. 54, n. 3, p. 507–554, 2005. Citado 4 vezes nas páginas 1, 3, 4 e 8.

RIGBY, R. A.; STASINOPOULOS, D. M. Using the box-cox t distribution in gamlss to model skewness and kurtosis. *Statistical Modelling*, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 3, p. 209–229, 2006. Citado na página 5.

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

RIGBY, R. A. et al. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. [S.l.]: CRC press, 2019. Citado 3 vezes nas páginas 1, 4 e 14.

RIPLEY, B. D. *Modern applied statistics with S*. [S.l.]: springer, 2002. Citado na página 9.

SCHNEIDER, L. et al. Model selection of nested and non-nested item response models using vuong tests. *Multivariate Behavioral Research*, Taylor & Francis, v. 55, n. 5, p. 664–684, 2020. Citado na página 5.

SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, JSTOR, p. 461–464, 1978. Citado na página 12.

STASINOPOULOS, D. M.; RIGBY, R. A. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, p. 1–46, 2008. Citado 2 vezes nas páginas 4 e 6.

STASINOPOULOS, D. M. et al. P-splines and gamlss: a powerful combination, with an application to zero-adjusted distributions. *Statistical Modelling*, SAGE Publications Sage India: New Delhi, India, v. 23, n. 5-6, p. 510–524, 2023. Citado na página 5.

STASINOPOULOS, M. D.; RIGBY, R. A.; BASTIANI, F. D. Gamlss: A distributional regression approach. *Statistical Modelling*, SAGE Publications Sage India: New Delhi, India, v. 18, n. 3-4, p. 248–273, 2018. Citado na página 4.

STASINOPOULOS, M. D. et al. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: CRC Press, 2017. Citado 7 vezes nas páginas 1, 4, 9, 10, 15, 16 e 17.

VONCKEN, L.; ALBERS, C. J.; TIMMERMAN, M. E. Model selection in continuous test norming with gamlss. *Assessment*, Sage Publications Sage CA: Los Angeles, CA, v. 26, n. 7, p. 1329–1346, 2019. Citado na página 5.

VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: journal of the Econometric Society*, JSTOR, p. 307–333, 1989. Citado 3 vezes nas páginas 2, 5 e 11.

WOOD, S. N. *Generalized additive models: an introduction with R*. [S.l.]: CRC press, 2017. Citado na página 3.