UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Luis Gonzaga Pinheiro Felix

Explorando seleção de variáveis explicativas no contexto dos modelos aditivos generalizados de locação, escala e forma

Recife

2023

Luis Gonzaga Pinheiro Felix

Explorando seleção de variáveis explicativas no contexto dos modelos aditivos generalizados de locação, escala e forma

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientadora: Profa. Dra. Fernanda De Bastiani

Coorientador: Prof. Dr. Daniel Matos de Carvalho

Recife

2023

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

F316e Felix, Luis Gonzaga Pinheiro
Explorando seleção de variáveis explicativas no contexto dos modelos aditivos generalizados de locação, escala e forma / Luis Gonzaga Pinheiro Felix. – 2023.
92 f.: fig., tab.

Orientadora: Fernanda De Bastiani.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2023.
Inclui referências.

1. Estatística aplicada. 2. Seleção de modelos. I. De Bastiani, Fernanda (orientadora). II. Título.

310 CDD (23. ed.) UFPE - CCEN 2023-38

LUÍS GONZAGA PINHEIRO FELIX

EXPLORANDO SELEÇÃO DE VARIÁVEIS EXPLICATIVAS NO CONTEXTO DOS MODELOS ADITIVOS GENERALIZADOS DE LOCAÇÃO, ESCALA E FORMA

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 17 de fevereiro de 2023.

BANCA EXAMINADORA

Profa. Dra. Fernanda De Bastiani
Orientadora, UFPE

Prof. Dr. Getúlio José Amorim do Amaral
Titular interno

Prof. Dr. Marcelo dos Santos
Examinador Externo à Instituição, IFBA

À minha família, em especial a MARIA DO SOCORRO PINHEIRO FELIX (in memoriam), minha mãe, por me apoiar em todos os momentos e está sempre ao meu lado.

AGRADECIMENTOS

A Deus, por tudo e tudo essa força superior, que sempre iluminou meus passos no caminho do bem e dos bons, sempre do meu lado.

A minha família, por está sempre comigo, apoiando e ajudando sempre em tudo.

A **Professora Drª Fernanda De Bastiani** que desde o início desta jornada sempre me orientou, mostrando caminhos e incentivando sempre para que mais esse passo fosse dado em minha vida.

Ao meu coorientador,**Professor Dr. Daniel Matos**, que foi meu colega de sala de aula, meu amigo e agora coorientador.

Aos meus professores do Departamento de Estatística em especial ao Professor Dr. Raydonal Ospina e a Professora Drª Maria do Carmo por estarem sempre a disposição nos momentos que precisei e ao Professor Dr. Francisco Cribari por todo conhecimento que transmitiu neste percurso.

Ao meu amigo, colega do curso e quase conterrâneo **André Medeiros**. Aos todos os amigos que conquistei durante o curso no departamento de Estatística, em especial os amigos José Jairo e Tatiane, Rodolfo Jordan, Lucas Miranda e Anny Kerol, Lucas David, Diego Rico, Charles Peixoto, Ranah, Cristini, Lucas Araujo, Larissa, Talita, Ana Bethy e Eduardo, Yuri Marti, Fernando, José Carlos, Luciano e a todos que estiveram comigo nesta caminhada.

A UFPE, instituição a qual devo toda minha formação acadêmica durante esta jornada. ‘ Aos meus amigos, companheiros de vida, que dividem comigo todo fardo e toda glória durante esta jornada. Obrigada!

RESUMO

A seleção de variáveis explicativas em modelos estatísticos é um problema atual e importante dentro da estatística e para o qual diferentes soluções já foram propostas para os diversos tipos de modelos. No caso específico dos modeloS aditivos generalizados de locação, escala e forma (GAMLSS), a seleção de variáveis explicativas é feita usando métodos *stepwise*. Na atual proposta de seleção de variáveis nos GAMLSS tem-se disponível duas estratégias conhecidas como estratégia A e estratégia B, sendo que ambas selecionam variáveis explicativas para modelar todos os parâmetros da distribuição, mas de forma diferente. Neste trabalho, estas metodologias foram descritas de forma detalhada e por meio de estudos de simulação, foram investigados e comparados métodos de seleção de variáveis, usando modelos com estruturas lineares e com estruturas não lineares, além do uso de funções de suavização para diferentes distribuições de probabilidade. Uma nova proposta de seleção de modelos, denominada de Estratégia C foi introduzida. A mesma é uma combinação da Estratégia B com a Estratégia A. Uma aplicação à dados reais ilustra a metodologia apresentada.

Palavras-chaves: critério de informação de Akaike; GAMLSS; P-splines; seleção de modelos; stepwise.

ABSTRACT

The selection of explanatory variables in statistical models is a current and important problem within statistics and for which different solutions have already been proposed for the various types of models. In the specific case of generalized additive models of location, scale and shape (GAMLSS), the selection of explanatory variables is done using *stepwise* methods. In the current proposal for variable selection in GAMLSS, two strategies known as strategy A and strategy B are available, both of which select explanatory variables to model all the parameters of the distribution, but in a different way. In this paper, these methodologies were described in detail and by means of simulation studies, variable selection methods were investigated and compared, using models with linear and with non-linear structures, as well as the use of smoothing functions for different probability distributions. A new proposal for model selection, called Strategy C, has been introduced. This is a combination of Strategy B and Strategy A. An application to real data illustrates the methodology presented.

Keywords: Akaike information criterion; GAMLSS; model selection; P-splines; stepwise.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os modelos estatísticos em geral são um conjunto de hipóteses sobre um conjunto de observações e/ou conjunto de dados, o que se caracteriza como uma metodologia extremamente vantajosa no sentido de interpretar e resumir informações mais importantes em um determinado estudo, além de ser bastante úteis para analisar relações entre variáveis aleatórias. Em um contexto mais simples, para falar de modelos estatísticos, podemos citar as relações lineares em que modelos normais lineares foram utilizadas ao longo de vários anos com o objetivo de descrever e explicar fenômenos aleatórios. Um exemplo são os modelos lineares, na qual estudos mostram que a suposição de linearidade entre a variável resposta e a variável explicativa em sua maioria não são adequadas e dessa forma sugerem mensurar o efeito dessas variáveis por meio de funções de suavização, flexibilizando a suposição de linearidade e com isso, permite-se também uma maior flexibilização na relação entre as variáveis. Atualmente, com a acelerada evolução tecnológica, tem-se expandido o leque de opções para a distribuição da resposta com novas classes de modelos.

Em situações práticas, os modelos estatísticos podem ser obtidos utilizando todas as variáveis de um conjunto de dados, contudo, há situações em que é comum observarmos muitas variáveis irrelevantes, com pouquíssima ou nenhuma correlação com a variável resposta, variáveis ruidosas ou não confiáveis. Com isso, se faz necessária a sua remoção para melhorar sua capacidade preditiva, reduzindo sua complexidade, melhorando suas propriedades estatísticas e definindo seus preditores mais rápidos e confiáveis.

Neste contexto, a seleção de variáveis explicativas se configura na estatística como um dos temas de grande relevância e que deve ser considerado no ajuste de modelos estatísticos. Assim, existe uma vasta quantidade de sugestões sobre como escolher o melhor modelo. Com isso, faz-se necessários o uso dos critérios de seleção de modelos, que são regras usadas para selecionar um modelo estatístico entre um conjunto de modelos candidatos, com base em dados observados. Por exemplo, Miller (1990) aborda sobre seleção de subconjuntos de variáveis explicativas, que mesmo com o passar dos anos, ainda permanece a necessidade de se avaliar a qualidade destas formas de seleção de variáveis explicativas na escolha do modelo.

Enfatizamos que nos modelos lineares normais de regressão, os métodos de seleção de modelos buscam selecionar variáveis explicativas para melhorar o ajuste da média da distribuição normal, ou seja, no ajuste de μ. No contexto dos modelos aditivos generalizados de locação,

escala e forma (*Generalized linear models for location, scale and shape* - GAMLSS - em inglês) desenvolvido por Rigby e Stasinopoulos (2005), quando se pensa em seleção do melhor modelo temos que selecionar principalmente variáveis explicativas para todos os parâmetros da distribuição, além da escolha da distribuição e da função de ligação. Atualmente, os critérios de seleção de modelos mais utilizados são o Critério de Informação de Akaike (em inglês *Akaike information criterion* - AIC) e critério de informação bayesiano (em inglês *Bayesian information criterion* - BIC) propostos por Akaike (1973) e Schwarz (1978b), respectivamente.

Stasinopoulos et al. (2017) apresentam os métodos de seleção de modelos utilizados pelo GAMLSS e sua implementação no R (R é um ambiente de *software* livre para computação estatística e gráfica), mas não detalham os procedimentos internos da função e não comparam entre os métodos de seleção disponíveis no pacote do **gamlss** e em outros pacotes do R. O estudo mais recente que aborda mais detalhes sobre seleção de variáveis explicativas no contexto do GAMLSS é apresentado em Ramires et al. (2021). Nesta pesquisa, os autores realizaram vários estudos de simulação para investigar a performance dos métodos de seleção de variáveis explicativas conhecido como Estratégia A e Estratégia B. Assim, o objetivo desta pesquisa dissertativa é explorar o comportamento das técnicas de seleção de variáveis explicativas, com foco no estudo da estratégia Estratégia A e Estratégia B, comparando-as no contexto dos modelos GAMLSS usando relações lineares e não lineares entre as variáveis explicativas e os parâmetros da distribuição.

Este estudo está organizado da seguinte maneira. O Capítulo 1 apresenta uma visão global da pesquisa, explicitando seus objetivos e a forma como a mesma está organizada. O Capítulo 2 apresenta uma breve descrição dos Modelos Lineares Generalizados - MLG, Modelos Aditivos Generalizados - GAM e dos Modelos GAMLSS, além de uma breve apresentação das principais funções suavizadoras. O Capítulo 3 descreve os principais procedimentos para seleção de variáveis explicativas, e descreve de forma rigorosa as Estratégia A e Estratégia B que foram utilizadas nos estudos de simulação. O Capítulo 4 discute como foram desenvolvidos os estudos de simulação, mostrando os resultados obtidos para as distribuições normal e Poisson inflacionada de zeros (em inglês *ZIP Zero-Inflated Poisson*) - ZIP, e compara os dois procedimentos de seleção de variáveis no contexto dos modelos GAMLSS. Também propõe uma seleção de modelos que combine as estratégias B e A. O Capítulo 5 compreende a aplicação que realizamos a um conjunto de dados reais com a finalidade de comparar as Estratégias A, Estratégia B e Estratégia C (que é uma combinação da Estratégia B com a Estratégia A). O Capítulo 6 apresenta as conclusões deste trabalho e sugestões para estudos posteriores.

2 CLASSES DE MODELOS ESTATÍSTICOS

Os modelos estatísticos constituem uma metodologia extremamente vantajosa no sentido de interpretar e resumir dados. Via de regra, são úteis para analisar relações entre variáveis aleatórias, além de serem fundamentais em investigações científicas e amplamente utilizados na caracterização da relação entre variáveis de interesse e seus efeitos. Dentre estas relações, as mais simples são as lineares, em que modelos normais foram utilizados ao longo de vários anos com o objetivo de descrever e explicar fenômenos aleatórios.

Apesar disso, na prática, a suposição de linearidade entre variáveis resposta e variáveis explicativas, em sua maioria, não são adequadas e a modelagem estatística sobre a suposição de normalidade das respostas podem ser altamente influenciadas por *outliers*, de modo que a aplicabilidade do modelo pode nos levar a conclusões errôneas.

Nesta conjuntura, alguns autores sugerem mensurar o efeito dessas variáveis por meio de funções arbitrárias, de modo que a suposição de linearidade seja relaxada permitindo ao modelo maior flexibilidade na relação entre as variáveis. Assim, nos últimos anos com o advento dos computadores, pensando em expandir o leque de opções para a distribuição da variável resposta, novos modelos foram criados, a exemplo os MLGs, os GAMs, os GAMLSS, dentre outros que ganharam destaque.

Para Zhang (1992), o impacto da seleção de variáveis em inferências estatísticas em modelos de regressão linear, usando em particular, o critério de erro de predição final generalizado de Shibata (1984), verificou entre outras coisas, que as inferências sobre os coeficientes de regressão são prejudicadas pelo procedimento de seleção de variáveis.

2.1 MODELOS LINEARES GENERALIZADOS - MLGS

Os MLGs foram propostos por Nelder e Wedderburn (1972) e surgiram como uma generalização da transformação de Box-Cox. A transformação de Box-Cox tinha o objetivo de estabilizar a variância e linearizar a relação entre o preditor linear e a variável resposta. Não sendo possível obter esses dois efeitos desejados, o MLG surgiu como uma opção para satisfazer esses dois efeitos, tendo como ideia básica expandir o leque de opções do modelo linear para a distribuição da variável resposta, permitindo que a mesma pertença a classe da família exponencial de distribuições.

2.1.1 Família Exponencial de distribuições

Sejam n variáveis aleatórias independentes $Y_1, \ldots, Y_n$, em que cada variável aleatória Y_i, com $i = 1, \ldots, n$, pertence a família exponencial de distribuição se sua função densidade de probabilidade for expressa da seguinte forma:

$$f(Y_i; \theta_i, \phi) = \exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y, \phi)\}, \quad \text{com} \quad i = 1, ..., n, \tag{2.1}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas.

Como $b(\cdot)$ é diferenciável e o suporte da distribuição não depende dos parâmetros, isto implica que é possível provar que a distribuição obedece as seguintes condições habituais de regularidade:

$$\mathbb{E}\left\{\frac{\partial \log(Y_i; \theta_i, \phi)}{\partial \theta_i}\right\} = 0$$

e

$$\mathbb{E}\left\{\frac{\partial^2 \log(Y_i; \theta_i, \phi)}{\partial \theta_i^2}\right\} = -\mathbb{E}\left[\left\{\frac{\partial \log(Y_i; \theta_i, \phi)}{\partial \theta_i}\right\}^2\right].$$

Para o modelo (2.1) valem as seguintes relações: $\mathbb{E}(Y_i) = \mu_i = b'(\theta_i)$, $Var(Y_i) = \phi^{-1}V(\mu_i)$, com $i = 1, \ldots, n$, sendo $\phi^{-1} > 0\,(\phi > 0)$ o parâmetro de dispersão e $V_i = V(\mu) = \frac{d\mu}{d\theta}$ a função de variância, que caracteriza a distribuição.

2.1.2 Especificação dos Modelos Lineares Generalizados - MLGs

Os MLGs são uma extensão do modelo linear clássico

$$\mathbf{y_i} = \mathbf{x}_i^\top \beta_i + \boldsymbol{\epsilon},$$

em que $\mathbf{x}$ é uma matriz $n \times p$, associada a um vetor $\beta_\mathbf{i} = (\beta_i, \cdots, \beta_p)^\top$ de parâmetros e $\boldsymbol{\epsilon}$ é um vetor de erros aleatórios com distribuição que se supõe normalidade. Quando se fala em extensão do modelo linear clássico, deve-se ver isto sobre dois aspectos, primeiro, que a distribuição considerada não precisa ser normal, e sim que pertença a família exponencial, e segundo, a relação entre a média e o vetor de covariáveis pode ser de qualquer função diferenciável.

É importante enfatizar que os MLGs permitem mais flexibilidade para a relação funcional entre a média da variável μ e o preditor linear η_i, ou seja, um MLG caracteriza-se pela seguinte

estrutura: uma componente aleatória, uma componente sistemática e uma função de ligação. Cordeiro e Demétrio (2008), enfatiza que uma decisão importante na escolha do MLG é definir os termos do trinômio: (i) a distribuição da variável resposta, sua escolha depende da natureza dos dados (discretos ou contínuos) e também do seu intervalo de variação (o conjunto dos reais, os reais positivos ou um intervalo real como (0,1)); (ii) a matriz do modelo $\mathbf{X} = \{x_{ij}\}$, com dimensão $n \times p$ deve ser supostamente de posto completo, pode representar o valor de uma covariável e sua forma representa, matematicamente o desenho do experimento; (iii) a função de ligação, esta depende do problema que está sendo modelado.

De modo geral, a formulação de um MLG deve levar em conta uma distribuição de probabilidade para a variável resposta, que deve pertencer à família exponencial de distribuições, as variáveis preditoras que podem ser qualitativas e/ou quantitativas e uma função de ligação que irá relacionar as componentes aleatórias com a sistemática do modelo.

Nestes termos, um MLG é definido por uma distribuição da família exponencial, uma estrutura linear e uma função de ligação. A função de ligação por sua vez depende do tipo de resposta e do estudo particular que se está a fazer e a mesma conecta a média da componente aleatória Y (esta pode ser de natureza contínua, discreta ou dicotômica) com o preditor linear, formado a partir de um conjunto de variáveis explicativas e representado pela componente sistemática. Assim, os MLGs podem ser definidos de acordo com as seguintes componentes:

a) Componente Aleatória: composta por uma variável aleatória **Y** com n observações independentes, um vetor de médias μ e uma distribuição de probabilidade pertencente a família exponencial.

b) Componente Sistemática: é estabelecido durante o planejamento do experimento e isto pode resultar em modelos de regressão, análise de variância, etc. Considerando a estrutura linear de um modelo de regressão, a parte sistemática é dada por $g(\mu) = \eta_i$, em que $\eta_{\mathbf{i}} = \mathbf{x}_i^\top \beta_{\mathbf{i}}$ é o preditor linear, $\beta_{\mathbf{i}} = (\beta_1, \cdots, \beta_p)^T$ é o vetor de parâmetros desconhecidos a serem estimados, $\mathbf{x}_i = (x_i, \cdots, x_n)^T$ representa os valores das variáveis explicativas (covariáveis) que acreditamos explicar parte da variabilidade inerente a $\mathbf{Y}$ e $g(\cdot)$ é uma função de ligação que deve ser monótona e diferenciável.

Nos modelos de regressão linear a relação entre o valor esperado da variável resposta e as variáveis explicativas é dada por $\mathbb{E}(Y_i) = \mu_i = \mathbf{x}_i^\top \beta_i$, porém nos MLG necessitamos de uma função de ligação que irá relacionar o componente aleatório ao componente sistemático e isto é feito através de uma função adequada a distribuição.

Para a especificação do modelo, os parâmetros θ_i da família exponencial canônica de distribuições não são de interesse direto (pois há um para cada observação) mas sim um conjunto menor de parâmetros $\beta_1, \beta_2, \ldots, \beta_p$, tais que uma combinação linear dos β_s seja igual à alguma função do valor esperado de Y_i.

Quanto a estimação dos parâmetros dos MLG, existem diversos métodos de estimação do vetor de parâmetros $\boldsymbol{\beta}$, o método de estimação por máxima verossimilhança é o mais usado em programas computacionais. Nelder e Wedderburn desenvolveram o algoritmo para estimar os parâmetros $\boldsymbol{\beta}$ em um MLG baseado no método Escore de Fisher, para mais detalhes ver Nelder e Wedderburn (1972).

2.2 MODELOS ADITIVOS GENERALIZADOS - GAM

Os modelos GAMs (*generalized additive models*), idealizado por Hastie e Tibshirani (1990) são uma extensão dos MLGs, com um preditor linear que envolve somas de funções suavizadoras das covariáveis. O modelo pode ser expresso da seguinte forma:

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots + f_p(x_{pi}, x_{p+1})$$

em que $\mu_i = \mathbb{E}(Y_i)$, com $g(\cdot)$ é uma função de ligação, Y_i é a variável resposta que segue uma distribuição pertencente à Família Exponencial, X_i^* é uma linha i do modelo paramétrico linear, θ^* é o vetor de parâmetro deste modelo e por fim as f_j são funções suaves das variáveis explicativas x_j do vetor e estão sujeitas a restrições de identificabilidade como $\sum f_j(x_{ij}) = 0$, para todo j.

Os GAMs tem flexibilidade quanto à especificação da dependência da resposta com as covariáveis. Ressalta-se que o uso de modelos com especificações suaves das covariáveis pode trazer mais informações do que o modelo paramétrico linear, no entanto, esta flexibilidade e conveniência traz duas situações importantes: a primeira, é a necessidade de representar esta função suavizadora de alguma maneira e a segunda é avaliar os ajustes destas funções.

As funções f_j são representadas por meio de bases *splines* de regressão, com medidas associadas de rugosidade da função que podem ser expressas como formas quadráticas nos coeficientes base. A estimação do modelo GAM, será realizada usando a estimativa de máxima verossimilhança penalizada, onde são utilizadas medidas de rugosidade para controlar os sobreajustes.

2.3 MODELOS ADITIVOS GENERALIZADOS PARA LOCAÇÃO, ESCALA E FORMA - GAMLSS

A classe dos modelos aditivos generalizados que consideram posição, escala e forma (GAMLSS) foi proposta por Rigby e Stasinopoulos (2005), e pode ser vista como uma extensão dos MLGs e GAMs de tal forma que supera algumas limitações destes dois modelos.

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

2.3.1 Definição

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

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.2}$$

em que $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \cdots, \theta_{nk})$ e $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \beta_{2k}, \cdots, \beta_{J_k'})$ são vetores de parâmetros de comprimento J_k'. A matriz do modelo $\boldsymbol{X}_k$ é conhecida de ordem $n \times J_k'$, $\boldsymbol{Z}_{jk}$ é uma matriz conhecida fixa de ordem $n \times q_{jk}$ e λ_{jk} é uma variável aleatória q_{jk}-dimensional, para mais detalhes veja (RIGBY; STASINOPOULOS, 2005).

Se $J_k = 0$, para $k = 1, 2, \ldots, p$ a expressão (2.2) é reduzida a um modelo paramétrico definido por

$$g_k(\theta_k) = \eta_k = X_k\beta_k.$$

Para Rigby e Stasinopoulos (2005), a expressão 2.2 possibilita que sejam incorporados diferentes tipos de efeitos aleatórios aditivos no modelo.

Usualmente, os dois primeiros parâmetros, θ_1 e θ_2 caracterizam a locação (μ) e a escala (σ). Para muitas famílias de distribuições, dois parâmetros de forma são suficientes: $\nu = \theta_3$ e $\tau =$

θ_4, para mais informações veja (FLORENCIO, 2010). De fato, a interpretação dos parâmetros depende de cada modelo. Costuma-se considerar distribuições até quatro parâmetros por ser a implementação disponível na literatura, no entanto, esta não é uma limitação teórica.

Assim, para quatro parâmetros tem-se o modelo

$$\begin{aligned} g_1(\mu) &= \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1} \\ g_2(\sigma) &= \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2} \\ g_3(\nu) &= \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3} \\ g_4(\tau) &= \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}. \end{aligned}$$

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

Para Stasinopoulos, Rigby et al. (2007), de modo geral, os parâmetros μ, σ, v e τ representam a locação, escala, assimetria e curtose, respectivamente, embora possam ser quaisquer outros parâmetros da distribuição. O pacote **gamlss** implementado em R permite escolher mais de 100 distribuições com até quatro parâmetros para a variável resposta, como por exemplo, Box-Cox-t, Gama, Pareto 2, Log-Normal, dentre outras.

2.3.2 Estimação dos parâmetros de um modelo GAMLSS

Na estrutura dos modelos GAMLSS é suposto um efeito aleatório com distribuição normal no preditor linear. O resultado da estimação utiliza uma matriz de suavização mediante um algoritmo de retroajuste (*backfitting*). Assim, assume-se no modelo 2.2 que γ_{jk}, à priori, possui distribuição normal independente com $\gamma_{jk} \sim N_{qjk}(0, \boldsymbol{G}_{jk}^{-1})$. sendo que $\boldsymbol{G}_{jk}^{-1}$ é a inversa generalizada de uma matriz simétrica de ordem $q_{jk} \times q_{jk}$, $\boldsymbol{G}_{jk} = \boldsymbol{G}_{jk}(\lambda_{jk})$. A matriz $\boldsymbol{G}_{jk}$ pode depender de um vetor de hiper-parâmetros λ_{jk} e caso seja singular λ_{jk} é dita ter uma função densidade proporcional a $\exp\left\{\frac{-1}{2}\gamma_{jk}^T \boldsymbol{G}_{jk}\gamma_{jk}\right\}$, para mais informação consulte (RIGBY; STASINOPOULOS, 2005).

Importante mencionar que dentro da estrutura dos modelos GAMLSS assume-se a independência entre os diferentes vetores aleatórios y_{jk}, mas se para um determinado k, dois ou

mais vetores de efeitos aleatórios não são independentes, os mesmos podem ser combinados em um único vetor aleatório com suas matrizes de projeto Z_{jk} em uma única matriz de projeto que satisfaça as condições de independência.

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

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\gamma_{jk}^{\top}\boldsymbol{G}_{jk}\gamma_{jk}$$

em que $l = \sum_{i=1}^{n}\log f(y_i|\theta^i)$ é a função log-verossimilhança, $\forall i = 1, 2, \cdots n$.

Para o modelo GAMLSS paramétrico a função log-verossimilhança penalizada (l_p) reduz-se a função log-verossimilhança (l) e os parâmetros β_k, com $k = 1,\ 2,\ 3$ e 4, são estimados maximizando a função l. Mais detalhes sobre como a função l_p é maximizada podem ser encontrados em Rigby e Stasinopoulos (2005). Utilizando o pacote **gamlss** em R, é possível avaliar as distribuições da variável dependente e dos diferentes termos aditivos do modelo na estrutura GAMLSS.

2.3.3 Preditor linear

Um detalhe importante quando se trabalha com análise de regressão é conhecer de forma profunda a estrutura do modelo na qual queremos usar para modelar os dados que temos disponível. Os GAMLSS permitem a seguinte estrutura: termos paramétricos, termos aditivos e combinação de termos.

a) Termos paramétricos: aqui, os preditores lineares η_k, com $k = 1, 2, \ldots, p$, são formados por componentes paramétricos $X_k\beta_k$ e aditivos $Z_{jk}Y_{jk}$, para $j = 1, 2. \ldots, j_{jk}$. Sendo que o componente paramétrico pode conter termos lineares e de interação , bem como fatores, polinômios e polinômios segmentados (com nós fixos) para as variáveis explicativas.

b) Termos aditivos: nos modelos GAMLSS com as componentes $Z_{jk}Y_{jk}$ é possivel modelar uma variedade de termos, como suavização e termos de efeito aleatório. Stasinopoulos et al. (2017) enfatiza que as variáveis explicativas podem afetar os parâmetros da distribuição especificada de diferentes maneiras. Os modelos GAMLSS permitem funções paramétricas lineares ou não lineares, ou funções de suavização não paramétricas de variáveis explicativas. O pacote **gamlss** permite diversos termos aditivos como: P-splines

(B-splines penalizadas), P-splines monótonas, P-splines cíclicos, P-splines de coeficiente variável, splines de suavização cúbica, polinômios fracionários, produto de tensor entre outros.

c) Combinação de termos: quaisquer combinações de termos paramétricos e aditivos podem ser combinadas (nos preditores de um ou mais dos parâmetros de localização, escala ou forma) para produzir termos ou modelos mais complexos.

2.3.4 Funções suavizadoras - splines

O termo spline tem sua origem em desenhos de engenharia, nos desenhos técnicos era usada para desenhar curvas que passam por pontos pré-determinados com o intuito de auxiliar na etapa de delineamento de objetos, tais como curvas de navios, peças de avião, etc. Sua formulação matemática foi desenvolvida oficialmente nos anos 60, porém existe registros que comprovam que já eram utilizado no século passado.

Pensar em spline consiste em interpolar um polinômio que passe pelos pontos do plano e por estes pontos é possível interpolar infinitos polinômios de diferentes graus, com o objetivo de se obter um melhor ajuste polinomial. Esta interpolação é feita por partes e os pontos de quebra (ruptura) são chamados de nós. Para Thomson e Emery (2014) as funções splines são muito utilizada em geociências, pois este método gera curvas mais suaves ao mesmo tempo tentando honrar o máximo de dados, mas mesmo assim, não é um interpolador exato.

O uso de splines tem suas vantagens e desvantagens, dentre as vantagens pode-se destacar a sua maior flexibilidade para o ajuste dos modelos se comparado ao modelo de regressão linear ou polinomial, a sua capacidade de modelar comportamentos atípicos dos dados, bem como a facilidade do ajuste quando determinados a quantidade e a posição dos nós.

Splines consiste basicamente em substituir o vetor x, por variáveis adicionais, que serão combinações de x e utilizar a aproximação linear nesse novo espaço. Considere a função $h(x)$, em que D representa o grau do polinômio, k o numero de nós e b_k o valor do nó, definida da seguinte forma:

$$h(x) = \sum_{j=1}^{D} \beta_{0j} x^j + \sum_{k=1}^{K} \sum_{j=0}^{D} \beta_{kj} (x - b_k)_+^j,$$

em que

$$(x - b_k)_+^j = \begin{cases} (x - b_k)^j, \text{ se } x \geq b_k \\ 0, \text{ se } x < b_k. \end{cases}$$

Após a definição de $h(x)$, o modelo passa a ser linear nessas novas variáveis geradas pelo vetor x e consequentemente, os métodos de estimação dos parãmetros poderão ser aplicados nesse novo espaço. Note que para cada intervalo entre os valores dos nós, tem-se um polinômio de grau D. Por exemplo, seja uma função de grau um $(D = 1)$ e com apenas um nó, então $h(x)$ poderá ser representada por:

$$(x - b_k)_+^j = \begin{cases} \beta_{00} + \beta_{01}x + \beta_{10} + \beta_{11}(x - b_1), se\ x\ \geq b_k \\ \beta_{00} + \beta_{01}, se\ x\ < b_k. \end{cases}$$

Perceba que em ambas as partes separadas pelos nós, temos uma função polinomial e como não possui nenhuma restrição, essa função é contínua, mas quando se analisa o valor dessa função no nó, veremos que a função não é contínua. Segundo Wood (2006), costuma-se escolher como representação de $h(x)$ funções capazes de aumentar a flexibilidade de $f(x)$ e, consequentemente, do modelo. Um exemplo disso são os polinômios, os mesmos são muito úteis quando o interesse reside nas propriedades de f na vizinhança de um ponto específico, no entanto, devido a sua natureza, tendem a distorcer a realidade em regiões remotas. Na estatística, como o interesse é o ajuste de curvas por meio de um único polinômio, acarreta problemas de multicolinearidade, isto é, alta correlação entre as variáveis explicativas. O que queremos é uma curva que se adéque a diferentes amostras vindas de uma mesma população e que seja capaz de captar adequadamente a variabilidade dos dados. Uma saída para lidar com essas dificuldades é ajustar polinômios de menor grau e por partes.

2.3.5 Funções suavizadoras - Splines Linear

É caracterizado como um conjunto de funções lineares com diferentes inclinações em cada intervalo definido pelos nós. A equação geral de um spline linear, em que k representa o numero de nós é dada por:

$$h(x) = \beta_{00} + \beta_{01} + \sum_{K=1}^{k} \beta_k (x - b_k)_+$$

em que

$$(x - b_k)_+ = \begin{cases} (x - b_1), \ se \ x \ \geq b_k \\ 0, se \ x \ < b_k. \end{cases}$$

2.3.6 Funções suavizadoras - B-splines

Um B-spline ou spline de base é uma função de spline com suporte mínimo em relação a um determinado grau, suavidade e partição de domínio. Qualquer função spline de determinado grau pode ser expressa como uma combinação linear de B-splines desse grau. Por exemplo, temos os B-splines cardinais que têm nós equidistantes uns dos outros.

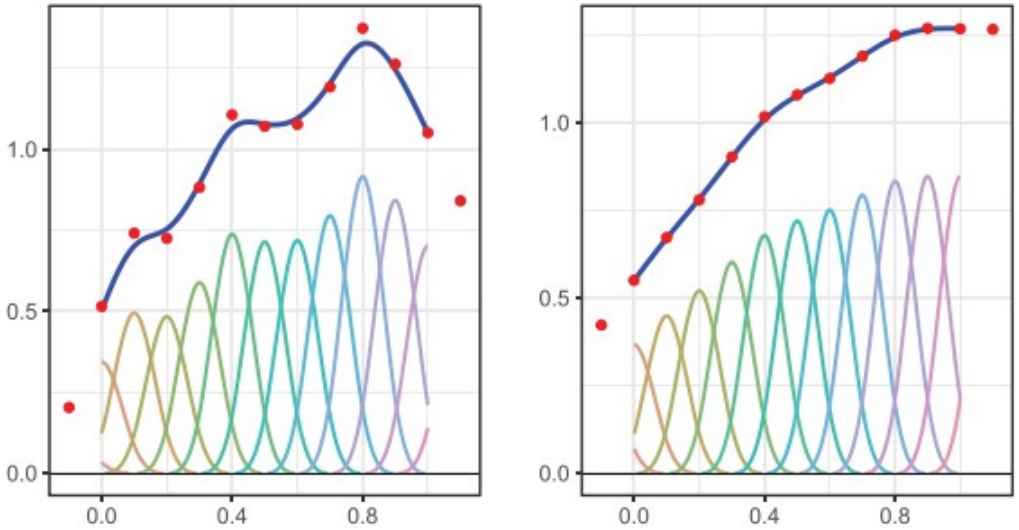
Os B-splines de ordem n são funções básicas para funções de spline da mesma ordem definidas nos mesmos nós, o que significa que todas as funções de spline possíveis podem ser criadas a partir de uma combinação linear de B-splines e há apenas uma combinação única para cada função de spline.

2.3.7 Funções suavizadoras - P-splines

O termo P-splines é uma combinação simples de duas ideias para a curva ajustada: a regressão sobre as funções de bases B-splines e uma penalidade de diferenças sobre os coeficientes da regressão dos B-splines, utilizando nós igualmente espaçados. Uma sugestão de Eilers e Marx (1996) é usar como penalização diferenças simples aplicadas diretamente aos parâmetros adjacentes dos B-splines. Os mesmos argumentam que essa combinação, a qual chamaram de P-splines, apresenta propriedades interessantes, tais como: não apresentam efeitos de contorno (ou fronteira); são uma extensão direta de modelos de regressão linear (generalizados), conservam momentos (média, variância) dos dados e têm curvas polinomiais ajustadas como limites, além de serem computacionalmente mais baratos.

Para Eilers e Marx (2021) P-splines são diferentes, os valores dos coeficientes estão próximos do valor da curva ajustada diretamente acima do pico da B-spline correspondente e a interpretação dos parâmetros é fácil: eles prevêem de perto a curva ajustada no centro do correspondente B-spline. Segue um exemplo, mostrado na Figura 1 e que ilustra bem o ajuste de uma curva por p-spline.

Figura 1 – Representação de curvas ajustadas por p-splines.



Fonte: (EILERS; MARX, 2021)

3 SELEÇÃO DE VARIÁVEIS EXPLICATIVAS

Na análise de regressão o comportamento da variável resposta pode ser explicado por outras variáveis que compõem a pesquisa. Incluir ou não uma ou mais variáveis vai depender do quanto estas variáveis são importantes para explicar a resposta e do tipo de modelo que o pesquisador quer construir. A necessidade de termos um modelo de regressão que melhor se ajuste aos nossos dados e consequentemente à situação problema, nos remete a necessidade de sabermos quais variáveis explicativas devem ser incluídas no mesmo. O que implica em escolher uma boa técnica de seleção de modelo e consequentemente um bom método de seleção de variáveis explicativas. Segundo Thompson (1989), os métodos analíticos passo a passo podem estar entre as práticas de pesquisa mais populares empregadas tanto na pesquisa substantiva quanto na de validade.

Como comumente empregados, esses métodos permitem a entrada de variáveis de previsão uma etapa por vez e a cada etapa também é considerada a remoção das variáveis inseridas anteriormente. Os métodos parecem ser empregados um tanto casualmente, especialmente em pesquisas de regressão e análise discriminante, embora também existam variantes quando outras técnicas são usadas.

Bayer e Cribari-Neto (2015) baseando-se no esquema de seleção de variáveis em dois passos, introduziram um esquema de seleção de variáveis para a regressão Beta em até k passos com o propósito de aumentar a proporção de vezes que o esquema de seleção de variáveis consegue encontrar um bom modelo (ou o "melhor modelo") para explicar as relações presentes entre as covariáveis e a variável resposta, sem aumentar de forma excessiva o custo computacional.

Já Efroymson (1960) propôs escolher as variáveis explicativas para um modelo de regressão múltipla a partir de um grupo de variáveis candidatas, passando por uma série de etapas automatizadas. Em cada etapa, as variáveis candidatas são avaliadas uma a uma, geralmente usando a estatística t para os coeficientes das variáveis que estão sendo considerada.

Existem atualmente vários procedimentos e critérios utilizados para a seleção de um subconjunto de variáveis regressoras para serem incorporadas no modelo. Dentre os vários métodos e procedimentos existentes na literatura, temos: seleção *best subset*, (seleção *stepwise*, seleção *backward*, seleção *forward*) (HOCKING, 1976), regressão ridge Hoerl e Kennard (1970), etc. Todos esses métodos buscam satisfazer a relação de compromisso entre o erro dentro da amostra e a parcimônia do modelo desejado.

A seguir será apresentado um breve resumo de alguns importantes métodos de seleção de variáveis explicativas que foram usados como base para os procedimentos aqui avaliados no contexto dos modelos GAMLSS.

3.1 MÉTODO *BACKWARD*

O método *backward* - passo atrás - caracteriza-se por incorporar inicialmente, todas as variáveis explicativas e percorrer etapas, nas quais uma variável por vez pode vir a ser eliminada. Caso em uma dada etapa, não ocorrer uma eliminação de alguma variável, o processo é então interrompido e as variáveis restantes definem o modelo final, veja (CHARNET et al., 1999).

A eliminação das variáveis numa dada etapa é feita da seguinte maneira: temos um modelo que denominaremos modelo completo da etapa e a partir daí investiga-se as contribuições individuais das variáveis a esse modelo. A variável explicativa de pior desempenho é eliminada, a não ser que esta atenda a um critério mínimo exigido. A escolha desse critério depende da classe de modelos que está sendo usada.

Segue abaixo, a sumarização dos passos de uma etapa deste procedimento, usando o teste F como critério para investigar as contribuições de cada uma das variáveis:

Passo 1: Ajuste o modelo de regressão com todas as suas k variáveis explicativas disponíveis;

Passo 2: Remova uma a uma as variáveis explicativas ao modelo, use um critério (teste F por exemplo) para avaliar o nivel de significância;

Passo 3: Seja $p_j = max(p_1, p_2, \ldots, p_k)$, em que $1 \leq j \leq k$, p indica o p-valor e α o nível de significância. Se $p_j > \alpha_c$, então a variável x_j correspondente é eliminada do modelo;

Passo 4: Os passos $\mathbf{1}$ à $\mathbf{3}$ são repetidos para o novo modelo (sem x_j e o processo continua até que $p_j = max(p) < \alpha_c$. Interrompa o processo e declare o modelo atual como selecionado.

3.2 MÉTODO *FORWARD*

A seleção para frente (passo *forward*) é uma técnica que começa sem variáveis na equação e adiciona uma variável por vez até que todas as variáveis estejam dentro ou até que um critério de parada seja satisfeito. A variável considerada para inclusão em qualquer etapa é aquela que produz o maior grau de liberdade entre aqueles elegíveis para inclusão. É importante enfatizar que o critério de parada a ser utilizado dependerá do tipo ou classe de modelo que está sendo

usado no momento.

Charnet et al. (1999), usando os modelos lineares, define o método de seleção *forward* da seguinte forma: dado o modelo reduzido e comparamos com modelos em que uma nova variável é acrescentada. Se há um modelo de melhor desempenho dentre os modelos com uma variável a mais e que atenda a um critério mínimo exigido, a correspondente variável é incorporada ao grupo de variáveis já escolhidas. Assim, enquanto em uma etapa do procedimento "passo atrás" comparamos vários modelos reduzidos com um único modelo completo, devido ao objetivo de eliminar uma variável, em uma etapa do procedimento "passo a frente", as comparações são feitas entre vários e um único modelo reduzido, devido ao objetivo de incorporar uma variável.

Segue abaixo, a sumarização dos passos de uma etapa deste procedimento:

Passo 1: Ajuste o modelo constante (sem covariáveis);

Passo 2: Adicione uma a uma as variáveis explicativas ao modelo, use um critério (teste F por exemplo) para avaliar o nível de significância;

Passo 4: Seja $p_j = min(p_1, p_2, \ldots, p_k)$, com $1 \leq j \leq k$ e p indica indicando o *p-valor* e α o nível de significância. Se $p_j < \alpha_c$, então a variável x_j correspondente é incorporada ao modelo;

Passo 4: Os passos $\mathbf{1}$ à $\mathbf{3}$ são repetidos para o novo modelo (com x_j e o processo continua até que $p_j = min(p) > \alpha_c$. Interrompa o processo e declare o modelo atual como selecionado.

3.3 MÉTODO *STEPWISE*

Existem vários métodos propostos na literatura para avaliar a eficácia de um modelo de regressão e uma das formas é adicionar ou eliminar suas covariáveis. Neste cenário, dos procedimentos e técnicas de seleção de variáveis explicativas citadas anteriormente, o método *stepwise*, é usado para selecionar quais variáveis que mais influenciam o conjunto de saída podendo, assim, diminuir o número de variáveis a compor o modelo de regressão.

O método *stepwise* é um desses métodos propostos que consiste em dois procedimentos: o *forward* e o *backward*. Esses procedimentos, geralmente como métodos conhecidos como métodos graduais, consistem em variações na mistura de duas ideias básicas chamadas de seleção direta e eliminação reversa, ver detalhes em Hocking (1976).

Alves, Lotufo e Lopes (2013) definem os procedimentos realizados no método *stepwise* da seguinte forma:

Passo 1: Escolhe-se a variável x_k que possui o maior coeficiente de correlação para entrar no modelo;

Passo 2: Uma variável x_i entra no modelo, se o coeficiente de correlação for maior que o anterior, x_i permanece no modelo, caso contrário, x_i sai do modelo;

Passo 3: x_i sai do modelo e se o coeficiente de correlação for menor que o anterior, x_i fica no modelo, caso contrário, x_i permanece fora do modelo. Este passo é repetido até que não tenha mais x_i para sair do modelo. Terminada esta etapa retorna-se ao **Passo 2** e este passo continua até que não tenham mais variáveis para entrar no modelo.

Por outro lado Thompson (1989) argumenta que um grande problema com a seleção *stepwise* é que uma otimização local obtida pela inclusão de variáveis uma a uma não é necessariamente uma otimização global. Por exemplo, selecionar uma quinta variável explicativa contingente às quatro variáveis já escolhidas não necessariamente seleciona as cinco variáveis que fornecem o maior R^2 possível. A técnica de regressão *stepwise* embora seja amplamente utilizadas em inúmeras pesquisas na área da estatística, se não for usada com muita cautela pode mostrar situações não reais e enganosas. Por exemplo, Smith (2018) fez uma série de simulações de Monte Carlo para demonstrar que a regressão *stepwise* é uma solução pobre para um excesso de variáveis e conclui que a mesma é menos eficaz quanto maior o número de variáveis explicativas potenciais. Na verdade, quanto maior o número de variáveis explicativas potenciais, mais provável é que a regressão *stepwise* seja enganosa.

Yamashita, Yamashita e Kamimura (2007) transformaram o método AIC usual no método *stepwise* AIC e comparam matematicamente o método *stepwise* AIC com outros métodos *stepwise* para seleção de variáveis, como Partial F, Partial Correlation e Semi-Partial Correlation em modelagem de regressão linear AIC e os outros métodos diferentes levam ao mesmo método que é o Partial F e concluíram que há mais razões para usar o método *stepwise* AIC do que os outros métodos *stepwise* para seleção de variáveis, uma vez que o método *stepwise* AIC é um método de seleção de modelo que pode ser facilmente gerenciado e amplamente estendido para modelos mais generalizados, como modelos lineares generalizados, modelos não lineares e dados não distribuídos normalmente.

Bayer e Cribari-Neto (2017) propuseram um esquema de seleção de variáveis em dois passos para o modelo de regressão beta duplo: assumindo que o parâmetro de dispersão é constante, ajustam-se todos os possíveis modelos para a média e seleciona-se o modelo de acordo com algum critério definido (menor AIC, por exemplo); assumindo que o modelo selecionado na etapa anterior é correto para o submodelo da média, ajustam-se todos os possíveis modelos

para o parâmetro de dispersão e seleciona-se o modelo de acordo com algum critério definido. E através de estudos de simulação de Monte Carlo, concluíram que o esquema de seleção de variáveis em dois passos obteve um desempenho semelhante ou superior, em comparação com a seleção dentre todos os modelos, na proporção de escolha do modelo verdadeiro para o modelo de regressão beta duplo.

3.4 CRITÉRIOS PARA AVALIAÇÃO E COMPARAÇÃO DE MODELOS ESTATÍSTICOS

A eficácia de um modelo de regressão está diretamente relacionada à sua capacidade preditiva e/ou explicativa em um conjunto de dados independentes. Ressalta-se que o modelo que melhor se ajusta a seus dados nem sempre é aquele que tem mais variáveis explicativas. Nesse contexto, um dos principais problemas enfrentados em muitas pesquisas é o de selecionar modelos parcimoniosos, ou seja, um modelo capaz de proporcionar um bom ajuste e com a menor quantidade possível de parâmetros, e isto está diretamente relacionado ao problema de seleção de variáveis explicativas, que atualmente, configuram-se em um dos problemas mais importantes quando se pensa em pesquisas de modelos de regressão, isso porque em geral, a inclusão indiscriminada de novas variáveis preditoras, apesar de reduzir o erro dentro da amostra, acaba aumentando o erro fora da amostra, mas a não inclusão de boas variáveis preditoras podem causar um problema sério que é a perda do poder de previsão do modelo fora da amostra.

Na análise de regressão um de seus principais objetivos é sem dúvida explicar a relação entre as variáveis em estudo da maneira mais simples possível, assim, quanto maior o número de parâmetros no modelo, menos graus de liberdade teremos para os resíduos, além disso, teremos menor precisão para as inferências. De forma análoga, quanto mais parametrizado o modelo, melhor será seu ajuste, contudo menor será seu poder de generalização, ou seja, um baixo poder preditivo.

Nesse contexto, para obtermos o melhor modelo, deve-se ter também um bom método ou critério de avaliação ou comparação de modelos, principalmente no processo de seleção de variáveis explicativas. Quando se pensa em seleção de variáveis explicativas na classe de modelos GAMLSS, é muito importante saber escolher o melhor critério de seleção de modelos, para que o metodo de seleção de variáveis seja o mais eficaz possível. Dentre os diversos critérios de seleção de modelos, nesta pesquisa faremos uma breve abordagem sobre os critérios: AIC, BIC e o Critério de Informação de Akaike Generalizado - GAIC.

3.4.1 Critério de Informação de Akaike - AIC

O AIC desenvolvido por Hirotugu Akaike, é originado da minimização da informação de Kullback-Leibler (KULLBACK; LEIBLER, 1951) e proposto em Akaike (1973). O AIC é um critério que permite ao pesquisador avaliar a qualidade do ajuste do modelo paramétrico e é estimado pelo método da máxima verossimilhança. O AIC é definido por:

$$AIC = -2\sum_{i=1}^{n} logf(x_i \mid \hat{\theta}) + 2p = -2logL(\hat{\theta}) + 2p,$$

em que $L(\hat{\theta})$ é a função de verossimilhança maximizada (calculada com base nos EMV's do parâmetros) e p é o número de parâmetros livres no modelo. O termo de penalidade $2p$ atua como uma compensação pelo viés na falta de ajuste quando os estimadores de máxima verossimilhança são usados.

3.4.2 Critério de Informação Bayesiano - BIC

Proposto por Schwarz (1978a), o BIC, é um critério de avaliação de modelos que é definido em termos da probabilidade a posteriori e parte do pressuposto da existência de um "modelo verdadeiro" que descreve a relação entre a variável dependente e as diversas variáveis explicativas entre os diversos modelos sob seleção. Este critério é baseado no cálculo do logaritmo da função de verossimilhança não sendo necessário a especificação de distribuições a priori, ou seja o critério é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados.

Seja $F(x_n, \mid \hat{\theta})$ um modelo estatístico estimado por máxima verossimilhança, então o critério BIC é definido por:

$$BIC = -2\sum_{i=1}^{n} \log f(x_i, \mid \hat{\theta}) + \kappa \log n = -2 \log L(\hat{\theta}) + \kappa \log n,$$

em que κ é o número de parãmetros a serem escolhidos e n é o numero de observações da amostra. O critério BIC penaliza mais fortemente a complexidade do modelo que o critério AIC ao substituir p por $\log(n)$ como fator de penalização.

3.4.3 Critério de Informação de Akaike Generalizado - GAIC

Selecionar modelos em GAMLSS implica em selecionar a melhor distribuição para a variável resposta, os preditores adequados para os parâmetros da distribuição selecionadas, das funções

de ligação e dos hiperparâmetros. Nesse contexto, o Critério GAIC, proposto por (VOUDOURIS et al., 2012), é dado por:

$$GAIC(\kappa) = -2L(\hat{\theta}) + (\kappa \times gl),$$

em que L é o logaritmo da função verossimilhança e gl são os graus efetivos de liberdade do modelo ajustado, κ é constante e é a penalidade para cada grau de liberdade utilizado. Para Paiva, Freire e Cecatti (2008) o critério GAIC leva em consideração o número de parâmetros e de graus de liberdade utilizados no modelo para penalizar os modelos mais complexos e evitar sobreajustes aos dados em amostras de grandes tamanhos.

3.5 SELEÇÃO DE VARIÁVEIS EXPLICATIVAS NO CONTEXTO DOS GAMLSS

Seja $\mathcal{M} = (\mathcal{D}, \mathcal{G}, \mathcal{T}, \mathcal{L})$ um modelo GAMLSS em que $\mathcal{D}$ é a componente que especifica a distribuição da variável resposta, $\mathcal{G}$ é a componente que especifica o conjunto das funções de ligação $(g_1, g_2, \ldots, g_p)$ para os parâmetros $(\theta_1, \theta_2, \ldots, \theta_p)$, $\mathcal{T}$ especifica o conjunto de termos preditores $(t_1, t_2, \ldots, t_p)$ para os preditores $(\eta_1, \eta_2, \cdots, \eta_p)$ e $\mathcal{L}$ é a componente que especifica os hiperparâmetros de suavização que determinam a quantidade de suavização nas funções S_{kj}.

A construção de um modelo GAMLSS para um determinado conjunto de dados é um processo que consiste em comparar modelos concorrentes usando diferentes combinações das componentes $\mathcal{M} = (\mathcal{D}, \mathcal{G}, \mathcal{T}, \mathcal{L})$ o que gera um número muito grande de possibilidades a serem avaliadas e testadas. Com isso, para encontrarmos um modelo GAMLSS apropriado para qualquer um novo conjunto de dados, torna-se necessário que todas as componentes acima devam sem bem especificadas de forma mais objetiva possível.

Avaliar um desempenho de um modelo estatístico é avaliar sua capacidade explicativa ou preditiva em um conjunto de dados independentes, assim, no geral se reconhece que modelos superajustados, ou seja, modelos que têm um desempenho excelente, porém quando utilizamos os dados de teste o resultado é ruim ou modelos subajustados, que são modelos que em um cenário o desempenho do modelo já é ruim no próprio treinamento, ou seja, o modelo não consegue encontrar relações entre as variáveis e o teste nem precisa acontecer.

Nesse estudo focaremos especificamente sobre a componente $\mathcal{T}$, que especifica a seleção de variáveis. Na classe de modelos GAMLSS assim como em outras classes de modelos, a seleção de variáveis explicativas é um tema importantíssimo para a estatística. Stasinopoulos,

Rigby et al. (2007) define os procedimentos para seleção de variáveis explicativas da seguinte forma: deixe X_k como um conjunto de variáveis explicativas a ser considerada para modelar o parâmetro θ_k na classe de modelos GAMLSS, onde $\theta_k = (\theta_1, \theta_2, \theta_3, \theta_4) = (\mu, \sigma, \nu, \tau)$. Aqui X_k pode conter fatores e termos quantitativos que podem ser inseridos no modelo como termos adicionais lineares ou suavizadores.

No contexto dos GAMLSS, existem cinco funções implementadas no software R que podem auxiliar na seleção de termos de variáveis explicativas. As duas primeiras são as funções `addterm()` e `dropterm()` que são funções genéricas com suas definições originais definidas no pacote **MASS** de (VENABLES; RIPLEY, 2002) e permitem a adição ou remoção de um termo em um modelo, respectivamente.

Essas duas funções são blocos de construção para a função `stepGAIC()` (ou seja, se baseia nos métodos `dropterm()` e `addterm()` aplicados aos objetos gamlss), adequadas para a seleção passo a passo de modelos. Ambas as funções realizam a seleção do modelo passo a passo usando um critério GAIC. O `stepGAIC()` realiza a seleção do modelo passo a passo usando GAIC e é baseado na função `stepAIC()` dada na biblioteca **MASS**. Duas funções principais a ser consideradas são as funções `stepGAICAll.A()` e `stepGAICAll.B()`, que são baseadas na função `stepGAIC()`, mas usam estratégias diferentes para selecionar um modelo final apropriado.

A `stepGAIC` é uma função que utiliza um algoritmo *stepwise* para realizar a seleção de modelo passo a passo usando o GAIC. Esta função é baseada na função `stepAIC()` da biblioteca **MASS** de Ripley (2002). É uma função que se baseia nos métodos `dropterm()` e `addterm()` aplicados aos objetos gamlss. Temos ainda os métodos `drop1()` e `add1()` que são equivalentes aos métodos `dropterm()` e `addterm()`, respectivamente, diferenciando-se apenas nos argumentos padrão diferentes.

Segundo Stasinopoulos et al. (2017) as estratégias A e B são estratégias para selecionar termos aditivos usando o GAIC, para todos os parâmetros de distribuição, assumindo uma distribuição de resposta particular. O **GAIC** que é uma função que calcula o GAIC para uma dada penalidade k para um objeto GAMLSS ajustado. Nesta pesquisa, para o estudo de simulação foram considerados as seguintes penalidades: $\kappa = 2$ (AIC), $\kappa = log(n)$ (BIC) e $\kappa = [(log(n) + 2)/2]$ (GAIC).

3.6 ESTRATÉGIA A

Segue uma descrição passo a passo de como funciona a função `stepGAICll.A()` implementada no pacote **gamlss** que trata da chamada Estratégia A.

Para cada modelo a ser ajustado, primeiramente deverá ser fixada a distribuição e suas funções de ligação a serem consideradas.

1. Use um procedimento de seleção GAIC progressivo para selecionar um modelo apropriado para todos os parâmetros da distribuição. No caso de uma distribuição de quatro parâmetros seria para μ, com σ, ν e τ tidas como constante, seguindo os seguintes passos:

i. Ajusta-se o modelo nulo (só com o intercepto) $y = \beta_0$ e obtem-se o valor do AIC.

ii. Usa-se o procedimento *forward* (passo a frente), considere o modelo ajustado em **i.** (modelo nulo) e para cada variável explicativa $X_1, X_2, \cdots, X_k$ não pertencente ao modelo em **i.**, adicione uma a uma as **k** variáveis gerando k modelos - e considere o modelo completo com a adição de cada uma destas variáveis adicionadas e em seguida obtenha o respectivo AIC.

iii. Denote por c_{min} o valor do menor AIC obtido dos k modelos gerados em **ii.**, ou seja $c_{min} = min(AIC_1, AIC_2, \cdots, AIC_k)$, com $k > 0$ e denote por c_0 o valor do AIC obtido para o modelo ajustado em **i.**.

a) Se $c_{min} < c_0$, então a variável X_j, com $j \leq k$ do modelo que tem $AIC_j = c_{min}$ é incorporada no modelo, ou seja, teremos $y = \beta_0 + \beta_j x_j$., em seguida volta-se ao passo i, iniciando nova etapa;

b) Se $c_{min} > c_0$, interrompe-se o processo e opta-se pelo modelo $y = \beta_0$ (modelo nulo).

2. Dado o modelo para μ obtido em 1 e com ν e τ como constantes, use o precedimento de seleção GAIC para selecionar um modelo apropriado para σ, usando o método *forward* de acordo com os passos **i.**, **ii.** e **iii.** de **1**.

3. Dados os modelos para μ e σ obtidos em **1** e **2** respectivamente e com τ como constante, use o procedimento de seleção GAIC para selecionar um modelo apropriado para ν usando o método *forward* de acordo com os passos **i.**, **ii.** e **iii.** de **1**.

4. Dados os modelos para μ, σ e ν, usar o procedimento *forward* para selecionar um modelo apropriado para τ, seguindo os passos **i.**, **ii.** e **iii.** de **1**.

5. Dados os modelos para μ, σ e τ, obtidos em **1**, **2** e **4** respectivamente, use procedimento de seleção *backward*, a partir do modelo para ν dado em **3**, para selecionar um modelo apropriado para ν seguindo os passos a seguir:

i. Considere o modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. ajustado para ν obtido em **3** com k variáveis.

ii. Usando o procedimento *backward* (passo atrás) e considerando o modelo ajustado em **i.** (modelo com k variáveis), retire uma a uma cada uma das covariáveis $X_1, X_2, \ldots, X_k$ e considere o modelo reduzido com $k-1$ variáveis, gerando $k-1$ modelos - dado a retirada de cada variável - e obtenha o respectivo AIC de cada um do $k-1$ modelos gerados.

iii. Denote por c_{min} o valor do menor AIC obtido com a geração dos $(k-1)$ modelos gerados em **ii.**, ou seja, $c_{min} = min(AIC_1, AIC_2, \ldots, AIC_k)$, com $k > 0$ e seja c_0 o valor do AIC obtido para o modelo completo ajustado em **i.**.

a) Se $c_{min} > c_0$ interrompe-se o processo e opta-se pelo modelo completo obtido nesta etapa;

b) Se $c_{min} < c_0$, então retira-se a variável x_j, com $j \leq k$ e cujo valor do $AIC_j = c_{min}$ e voltamos ao passo **i.**, iniciando nova etapa em que o modelo completo terá $(k-1)$ variável regressoras, ou seja, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1}$, dado a eliminação da variável x_j cujo $AIC = c_{min}$.

iv. Os passos **i.** e **iii.** são repetidos para o novo modelo (quando x_j é retirada) e o processo continua até que $c_j = min(AIC) < AIC_0$. Neste caso, o processo é interrompido e o modelo atual será o selecionado.

6 Dados os modelos para μ, ν e τ, obtidos em **1**, **5**, **4**, respectivamente, use procedimento de seleção *backward*, a partir do modelo para σ dado em **2**, para selecionar um modelo apropriado para σ seguindo os passos dado em **5**.

7 Dados os modelos para σ, ν e τ, obtidos em **6**, **5** e **4**, respectivamente, use procedimento de seleção *backward*, a partir do modelo para μ dado em **1**, para selecionar um modelo apropriado para μ seguindo os passos dado em **5** e o processo é interrompido.

3.7 ESTRATÉGIA B

O procedimento de seleção de variáveis explicativa `stepGAICAll.B()`, que é usualmente conhecido como estratégia B usa o mesmo procedimento da função stepGAIC(), utilizados no procedimento `stepGAICAll.A()`, porem o que diferencia as dois procedimentos é que o `stepGAICAll.B()` seleciona cada termo no escopo e ajusta a todos os parâmetros da distribuição, ou seja, estratégia B força que um mesmo conjunto de variáveis explicativas selecionadas componha os preditores dos diferentes parâmetros do modelo.

Segue uma descrição passo a passo de como funciona a função `stepGAICll.B()`, Estratégia B implementada no pacote **gamlss**.

1 Ajustar o modelo reduzido (apenas com o intercepto) e obtenha o AIC.

2 Considere o modelo ajustado em **1** e use o procedimento *Forward* (passo a frente), para cada uma das variáveis $X_1, X_2, \ldots, X_k$ não pertencente ao modelo do passo **1** adicione uma a uma cada variável, gerando k modelos - com a adição desta variável extra - e obtenha o valor do AIC. Denote por b_{min} o valor do menor AIC obtido dos k modelos gerados em **2** ou seja $b_{min} = min(AIC_1, AIC_2, \ldots, AIC_K)$, com $k > 0$,.

Seja b_{01} o valor do AIC obtido para o modelo ajustado antes da inclusão da variável, então:

a) Se $b_{min} < b_{01}$, passar ao passo **3** com o modelo completo ajustado do passo **1** mais a variável do modelo cujo AIC é igual a b_{min}.

b) Se $b_{min} > b_{01}$, passar para o passo **3**, com modelo ajustado igual ao modelo do passo **1** ou encerrar o processo se no passo **5** da etapa anterior, se nenhuma variável tiver sido eliminada.

3 Ajustar o modelo completo (incluindo a variável adicionada no passo **2** e obter o valor do AIC.

4 Para cada uma das variáveis do modelo completo do passo **3**, considerar o modelo reduzido - retirando esta variável - e obter o AIC.

5 Denote por b_{min} o valor do maior AIC obtido dos modelos obtidos em **4** ou seja $b_{min} = min(AIC_1, AIC_2, \cdots, AIC_K)$, com $k > 0$,, Seja b_{02} o valor do AIC obtido para o modelo ajustado antes da inclusão da variável. Então:

Se $b_{min} > b_{02}$, não eliminar nenhuma variável e voltar ao passo **1**, iniciando nova etapa com modelo reduzido com k variáveis ou encerrar o processo se no passo **2** nenhuma variável tiver sido anexada.

Se $b_{min} < b_{01}$, eliminar a variável cujo AIC é igual ao b_{min} e voltar ao passo **1**, iniciando nova etapa com modelo reduzido com $(k-1)$ variáveis.

4 ESTUDO DE SIMULAÇÃO

Neste capítulo, é apresentado e discutido estudos de simulação usando o método de Monte Carlo com o objetivo de comparar os procedimentos para a seleção de variáveis explicativas em modelos GAMLSS abordados em 2.3. Para isto, usamos como suporte computacional o software R, especificamente o pacote `gamlss` e seus pacotes auxiliares desenvolvidos por (STASINOPOULOS; RIGBY et al., 2007).

O pacote `gamlss` foi elaborado baseando-se em (RIGBY; STASINOPOULOS, 2005), conforme descrito em Stasinopoulos, Rigby et al. (2007) e usa dois diferentes algoritmos para a obtenção das estimativas dos parâmetros do modelo, um deles denominado RS e o outro CG (RIGBY; STASINOPOULOS, 2005).O algoritmo RS costuma ser mais rápido em termos computacionais pois não necessita da informação de derivadas cruzadas.

Para o referido estudo de simulação, tomou-se como base o esquema proposto por Ramires et al. (2021), ou seja, foram consideradas as distribuições:

Distribuição Normal (N), $i.e.\ Y \sim N(\mu, \sigma)$, com $-\infty < y < \infty$, $-\infty < \mu < \infty$ (a média μ) e $\sigma > 0$ o parâmetro de dispersão;

Distribuição ZIP, $i.e.\ Y \sim ZIP(\mu, \sigma)$, com $y = \{0, 1, 2, \cdots\}$, e $\mu < 0$ é a média e $0 < \sigma < 1$ é a probabilidade exata de $Y = 0$. O modelo Poisson Inflado de Zeros – ZIP é um dos mais usados entre os modelos para dados de contagem. Ele é utilizado quando observamos em uma distribuição discreta de Poisson com maior quantidade de observações iguais a zero que o modelo permite.

Quanto a função de ligação usada, Stasinopoulos et al. (2017) e De Bastiani et al. (2018), sugerem que a escolha deve ser com base no suporte definido para cada parâmetro θ_k, ou seja, se o parâmetro é definido no conjunto dos números reais, a função de ligação utilizada deve ser a identidade, se o suporte de θ_k está definido no conjunto dos números reais positivos, a função de ligação é a logarítmica, se o suporte for no intervalo unitário, a função de ligação deverá ser a logit.

Este estudo foi desenvolvido levando em consideração primeiramente a estrutura do modelo de regressão, a saber: modelos de regressão com uma estrutura linear e modelos de regressão com uma estrutura não linear. Para todos os cenários, foram geradas 8 (oito) diferentes covariáveis para o processo de criação e seleção dos modelos, conforme descrito em Ramires et al. (2021), distribuídas da seguinte forma:

- Variáveis binárias: $X_1 \sim X_2 \sim X_3 \sim X_4 \sim Bernoulli(0,5)$
- Variáveis contínuas: $X_5 \sim X_6 \sim X_7 \sim X_8 \sim U(0,1)(ouseja, umadistribuição uniforme)$
- Foram considerados diferentes valores para seus coeficientes associados as variáveis X_5 e X_6, tanto para a estrutura do parâmetro μ como para a estrutura do parâmetro σ, mantendo todos os outros coeficientes constantes.
- As variáveis X_4 e X_8 foram consideradas como variáveis de ruído, ou seja, embora as mesmas estejam incluídas no procedimento de seleção de modelo, não foram consideradas no processo de geração de dados.
- Foram avaliados diferentes valores de inclinação do modelo.

As principais diferenças do estudo apresentado nesta dissertação em relação ao que foi apresentado por Ramires et al. (2021), é que aqui foi considerado as distribuições Normal e ZIP com tamanhos amostrais 150, 300 e 500 e com 1000 réplicas de Monte Carlo para avaliar e comparar o desempenho das Estratégias A e B. Além disso, uma nova proposta considerada foi a avaliação de uma Estratégia C, como sendo a combinação entre a Estratégia B e A, utilizando três penalidades diferentes. ou seja, os critérios AIC, BIC e GAIC.

Como indicador para avaliação dos processos de seleção de variáveis, foram considerados para cada replica do método de simulação de Monte Carlo, se cada um dos procedimentos de seleção de variáveis explicativas Estratégias A e B, selecionam corretamente ou não as variáveis que de fato deveriam ser selecionadas e ao final de todas as réplicas calcula-se o percentual de vezes que cada estratégia (A ou B) selecionou corretamente as mesmas, considerando diferentes valores de k. Além disso, é retornado um resumo dos p-valores para cada variável selecionada em cada réplica.

4.1 SELEÇÃO DE VARIÁVEIS EXPLICATIVAS USANDO AS ESTRATÉGIA A E ESTRATÉGIA B, EM MODELOS COM ESTRUTURA LINEAR

Para este cenário os dados simulados foram consideradas apenas relações lineares entre covariáveis para cada um dos parâmetros da distribuição, além disso, foi utilizado o esquema proposto por Ramires et al. (2021), que dentre as oito covariáveis, uma variável discreta (X_3) e uma variável continua (X_7) afetaram simultaneamente os parâmetros μ e σ, ou seja, as

variáveis x_3 e x_7 foram inseridas no submodelos para ambos os parâmetros, as variáveis X_1 e X_5 foram consideradas apenas no processo de geração para μ e as variáveis X_2 e X_6 para o parâmetro σ. Com relação aos tamanhos amostrais, foram considerados $n = 150$, 300 e 500, com 1000 réplicas de Monte Carlo.

4.1.1 Resultados do estudo de simulação para selecionar variáveis explicativas com dados normais e usando Estratégia A e a Estratégia B.

Primeiramente consideramos a variável aleatória $Y \sim N(\mu, \sigma)$ e seguimos a estrutura de regressão descrita por Ramires et al. (2021) dada por:

$$\mu = \beta_{01} + \beta_{11}x_1 + \beta_{31}x_3 + \beta_{51}x_5 + \beta_{71}x_7$$
$$\sigma = \exp[\beta_{02} + \beta_{22}x_2 + \beta_{32}x_3 + \beta_{62}x_6 + \beta_{72}x_7].$$

Com os seguintes valores verdadeiros dos parâmetros nos processos de geração de dados:

$$\mu = 40 + 5x_1 - 3x_3 + 2,5x_5 - 3x_7 \tag{4.1}$$
$$\sigma = exp[1,6 + 0,6x_2 - 0,35x_3 + 0,03x_6 - 0,02x_7] \tag{4.2}$$

As Tabelas 1, estão presentes os valores dos porcentuais das variáveis explicativas selecionadas corretamente, usadas no modelo 4.1 para diferentes tamanhos de amostra (150, 300 e 500) usando três tipos deferentes de penalidades (ou seja, diferentes valores de k). De acordo com resultados obtidos para a Estrategia A observa-se que para o parâmetro μ, as variáveis x_1 e x_3 obtiveram ótimos índices próximos ou igual a 100% de especificação correta, ou seja, foram selecionadas corretamente com um excelente percentual, para os 3 tamanhos de amostras considerados (150, 300 e 500) e para as três penalizações utilizadas.

No entanto, quando foi observado os percentuais obtidos para as variáveis x_5 e x_7, os resultados da Tabela 1 nos mostram que para o parâmetro μ as mesmas obtiveram percentuais bem menores, principalmente para tamanhos da amostrais menores, sendo que os piores resultados foram obtidos para $\kappa = log(n)$ (critérios BIC). Segundo Ramires et al. (2021), uma possível explicação para isso é o efeito de covariáveis contínuas no logaritmo da função de verossimilhança que é menor do que o exercido por uma variável categórica.

Para o parâmetro σ, de acordo com a Tabela 1 a Estratégia A obteve baixos percentuais, principalmente para as variáveis x_6 e x_7 quando o tamanho amostral cresce, chegando a menos de 2% quando $\kappa = log(n)$ e próximo de 20% para as demais penalidade utilizadas.

Mas observando as variáveis x_2 e x_3, percebe-se que a Estratégia A obteve ótimos percentuais de acertos, chegando muito próximo ou igual a 100%.

As variáveis explicativas x_2, x_4, x_6 e x_8, quando se considera o parâmetro μ e as variáveis explicativas x_1, x_4, x_5 e x_8, para o parâmetro σ, como era o esperado, obtiveram os valores percentuais muito baixos, ou seja, a Estratégia A obteve baixos percentuais para as variáveis de ruído.

Tabela 1 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo com estrutura linear normal, e diferentes valores para κ

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	$100,00$	$11,90$	$90,10$	$8,20$	$43,40$	$8,90$	$56,40$	$9,50$
300	μ	$100,00$	$11,10$	$99,60$	$8,00$	$67,50$	$8,30$	$82,50$	$8,50$
500	μ	$100,00$	$11,00$	$100,00$	$7,80$	$86,60$	$9,70$	$95,30$	$9,80$
150	σ	$17,30$	$100,00$	$90,50$	$15,60$	$16,90$	$19,30$	$17,20$	$16,50$
300	σ	$15,70$	$100,00$	$99,20$	$16,10$	$16,70$	$19,70$	$17,50$	$14,90$
500	σ	$17,40$	$100,00$	$100,00$	$16,60$	$17,00$	$17,20$	$14,60$	$16,10$
				Para $\kappa = log(n)$, (Critério BIC)					
150	μ	$99,40$	$2,10$	$71,40$	$1,00$	$16,40$	$1,40$	$25,60$	$0,70$
300	μ	$100,00$	$0,80$	$94,50$	$0,60$	$31,40$	$0,60$	$45,50$	$1,10$
500	μ	$100,00$	$1,00$	$99,80$	$0,50$	$49,70$	$0,50$	$69,90$	$0,50$
150	σ	$3,20$	$99,80$	$70,40$	$2,70$	$3,20$	$4,20$	$2,70$	$3,40$
300	σ	$2,40$	$100,00$	$95,20$	$1,70$	$2,20$	$2,40$	$2,00$	$2,50$
500	σ	$1,10$	$100,00$	$99,80$	$2,00$	$1,00$	$1,30$	$1,50$	$1,80$
				Para $\kappa = [2 + log(n)]/2$					
150	μ	$99,90$	$4,50$	$81,50$	$3,00$	$27,10$	$3,20$	$38,70$	$2,30$
300	μ	$100,00$	$2,80$	$97,60$	$1,80$	$47,50$	$1,70$	$64,90$	$2,60$
500	μ	$100,00$	$2,70$	$99,90$	$1,50$	$68,10$	$2,30$	$83,70$	$2,30$
150	σ	$7,20$	$100,00$	$81,00$	$6,90$	$7,20$	$9,00$	$7,50$	$7,30$
300	σ	$5,30$	$100,00$	$97,80$	$5,30$	$5,60$	$6,60$	$5,80$	$5,60$
500	σ	$4,90$	$100,00$	$99,90$	$5,80$	$4,40$	$4,70$	$4,70$	$4,50$

Fonte: Autoria própria (2023)

Ainda usando modelo 4.1, foi realizado simulação de Monte Carlo usando o procedimento de seleção de variáveis Estratégia B e os resultado são apresentados na Tabela 2 que trás os percentuais de acertos do procedimento de seleção da Estratégia B para os parâmetro μ e σ.

Foi observado que as variáveis x_2 e x_3, para todos os valores de κ considerados, obtiveram altos índices de especificação correta, muito próximos ou iguais a 100%, principalmente quando o tamanho da amostra é aumentado, enquanto que as variáveis x_5 e x_7 obtiveram percentuais de especificação correta muito baixo (menos de 50%), quando na realidade se esperaria o índice alto, visto que as mesmas fazem parte do modelo inicial. Mesmos com com baixos percentuais corretos, não podemos afirmar que é uma falha do procedimento. Uma justificativa pode ser os baixos valores para seus respectivos coeficientes que podem terem afetados diretamente.

Aqui foi observado um problema para a Estratégia B, e se repetirá em todos os cenários, a saber: por definição, a Estratégia B usa o procedimento stepGAIC utilizando os passos para frente e pra trás para selecionar as variáveis o que implica obrigatoriamente a inclusão desta variável selecionada para todos os parâmetros, com isso se configura um problema, pois variáveis de ruídos obtiveram excelentes índices como mostra a Tabela 2 ao selecionar as variáveis x_2 para o submodelo para μ e a variável x_1 para o submodelo para σ.

Assim, foi possível perceber que a Estratégia B obteve um auto índice de seleção correta para algumas variáveis, contrariando o que era esperado (baixo índice).

Quando comparados os resultados considerando os três tipos deferentes de penalidades usadas, percebe-se que com $\kappa = 2$ (AIC) a estratégia B seleciona melhor, ou seja, os percentuais de especificação correta são bem melhores, principalmente quando o tamanho da amostra aumenta.

Ainda de acordo com a Tabela 2, onde estão presentes os valores obtidos para o parâmetro σ, para a Estratégia B os percentuais especificação correta para as variáveis x_2 e x_3 são muito satisfatório, chegando a 100% quando o tamanho da amostra é aumentado e muito próximo a isso para tamanhos amostras menores, isso para os três valores de κ considerados. O mesmo não acontece com as variáveis x_6 e x_7, que fora do esperado, obtiveram baixos valores, menos de 50%, para todo os valores de κ aqui considerados.

Um resultado ruim para a Estratégia B é o altíssimo índice de especificação correta para algumas variáveis de ruído, principalmente a variável x_1, quando na realidade deveria ser o menor valor possível.

Tabela 2 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo com estrutura linear normal, e diferentes valores para κ.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	100,00	99,90	99,10	17,80	52,30	18,30	63,30	19,10
300	μ	100,00	100,00	100,00	13,40	74,40	15,90	86,90	14,90
500	μ	100,00	100,00	100,00	14,10	91,90	15,10	97,80	15,10
150	σ	100,00	99,90	99,10	17,80	52,30	18,30	63,30	19,10
300	σ	100,00	100,00	100,00	13,40	74,40	15,90	86,90	14,90
500	σ	100,00	100,00	100,00	14,10	91,90	15,10	97,80	15,10
				Para $\kappa = log(n)$, (Critério BIC)					
150	μ	99,40	96,20	89,00	0,80	9,30	0,70	19,60	1,20
300	μ	100,00	99,90	99,80	0,40	24,30	0,20	39,90	0,30
500	μ	100,00	100,00	100,00	0,70	41,90	0,30	64,90	0,40
150	σ	99,40	96,20	89,00	0,80	9,30	0,70	19,60	1,20
300	σ	100,00	99,90	99,80	0,40	24,30	0,20	39,90	0,30
500	σ	100,00	100,00	100,00	0,70	41,90	0,30	64,90	0,40
				Para $\kappa = [2 + log(n)]/2$.					
150	μ	100,00	99,30	96,50	4.70	25,60	4,60	39,00	4,40
300	μ	100,00	100,00	100,00	2,10	48,20	2,30	65,20	3,00
500	μ	100,00	100,00	100,00	2,50	69,10	2,30	87,20	1,70
150	σ	100,00	99,30	96,50	4,70	25,60	4,60	39,00	4,40
300	σ	100,00	100,00	100,00	2,10	48,20	2,30	65,20	3,00
500	σ	100,00	100,00	100,00	2,50	69,10	2,30	87,20	1,70

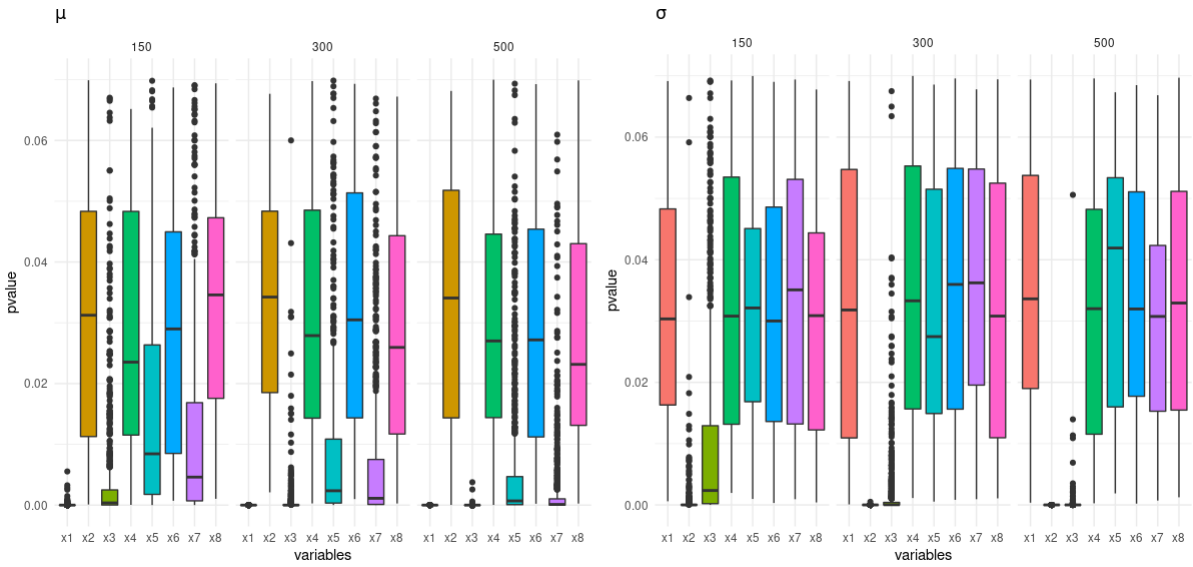
Fonte: Autoria própria (2023)

De acordo com as Tabelas 1 e 2 onde estão presente os percentuais de acertos dos procedimentos de seleção de variáveis Estratégias A e B, usando três tipos de penalidades diferentes, ou seja, três valores para κ, em um modelo normal, a conclusão foi que a `Estrategia A`, obteve melhores resultados.

Nos Boxplot apresentados nas Figuras 2 e 3 estão presentes os p-valores para o critério AIC ($\kappa = 2$) usando o procedimento de seleção de variáveis Estratégia A e Estratégia B respectivamente. Com os resultados obtidos é possível perceber que os p-valores tem maior dispersão para as variáveis de ruido, ou seja, que não estão inclusas no modelo, enquanto

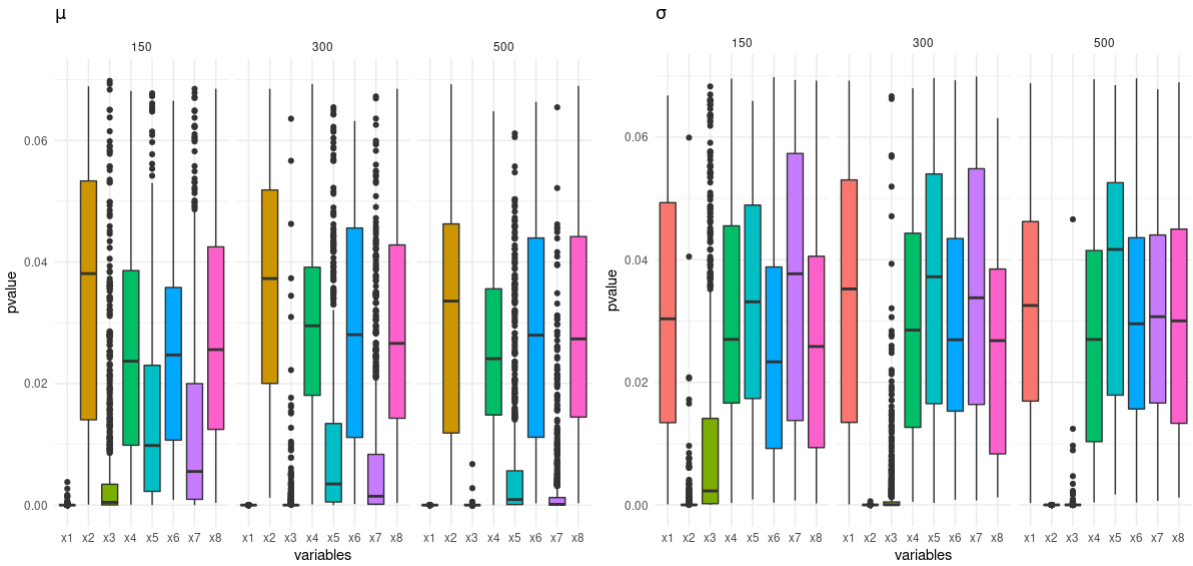
que para as variáveis autênticas (inclusas no modelos) os p-valores são bem menores e muito pouco disperso. Porém, para as variáveis inclusas no modelo, os Gráficos 2 e 3 apresentaram um numero bem grande de valores discrepantes, principalmente para o submodelo para μ para tamanhos amostrais pequenos.

Figura 2 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo normal com estrutura linear, com o critério AIC.



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 3 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo normal com estrutura linear, com o critério AIC.

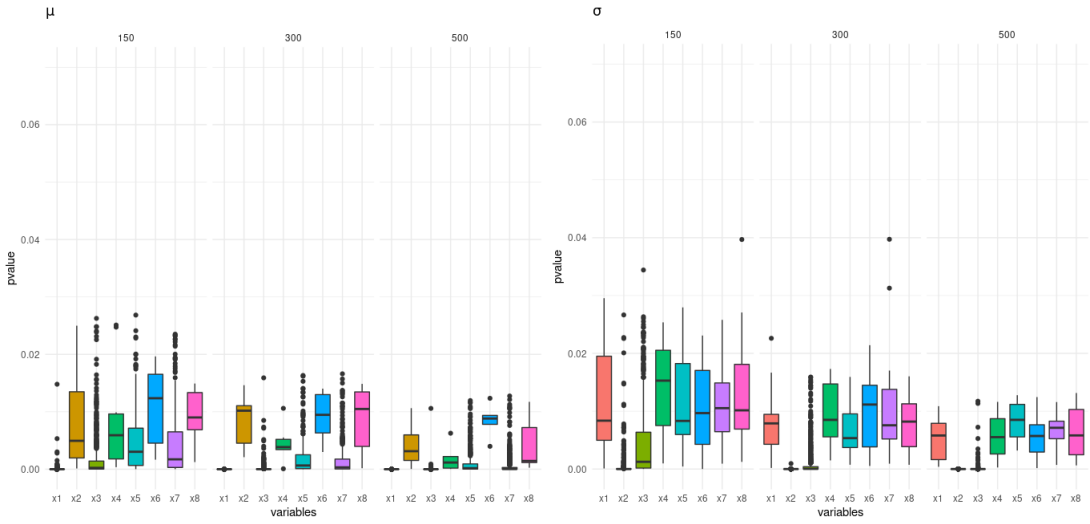


Fonte: Adaptado de (RAMIRES et al., 2021)

Os Boxplot apresentado nas Figuras 4 e 5 foi possível perceber que para a Estratégia B

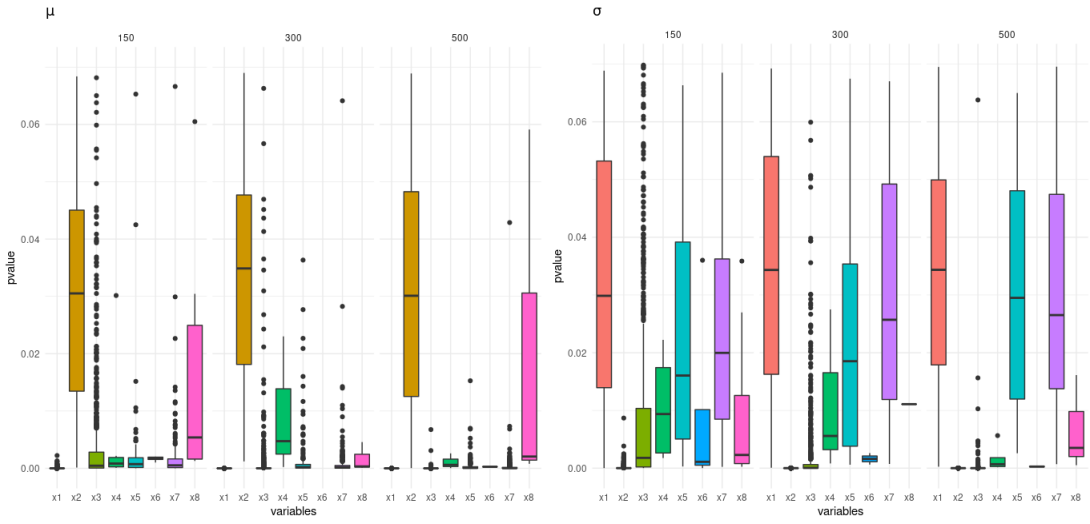
uma situação análoga ao critério AIC, ou seja, os valores dos p-valores estão bem dispersos para aquelas variáveis que não estão no modelo.

Figura 4 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo normal com estrutura linear, com o critério BIC.



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 5 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo normal com estrutura linear, com o critério BIC.

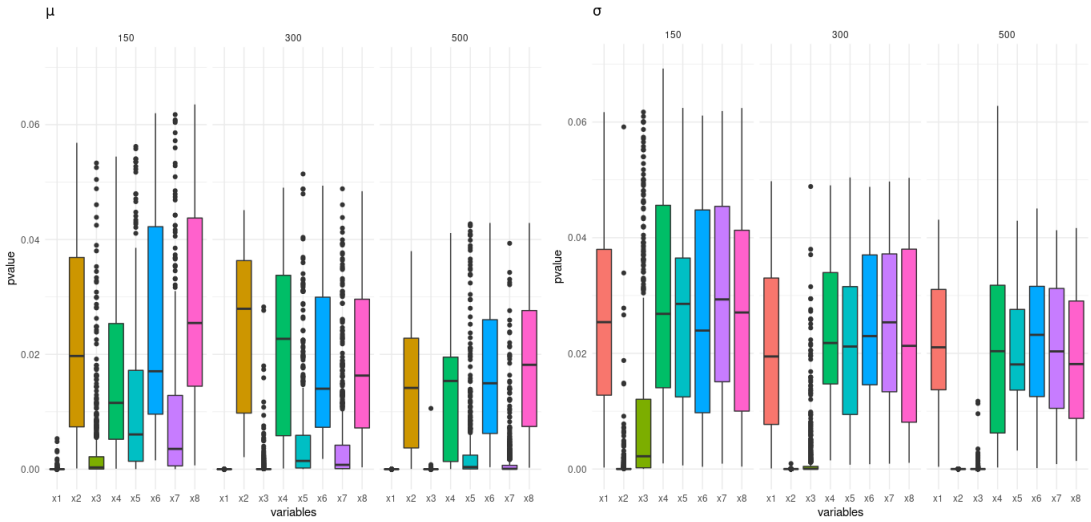


Fonte: Adaptado de (RAMIRES et al., 2021)

Os Boxplot 6 e 7 apresentam os p-valores para $\kappa = [(log(n)+2)/2]$ na seleção de variáveis explicativas usando as estratégias A e B. Um ponto de atenção que deve ser observado é a alta dispersão dos p-valores para as variáveis que não estão no modelo e os baixos valores dos p-valores e pouca dispersão para aquelas variáveis que estão no modelo. É importante observar

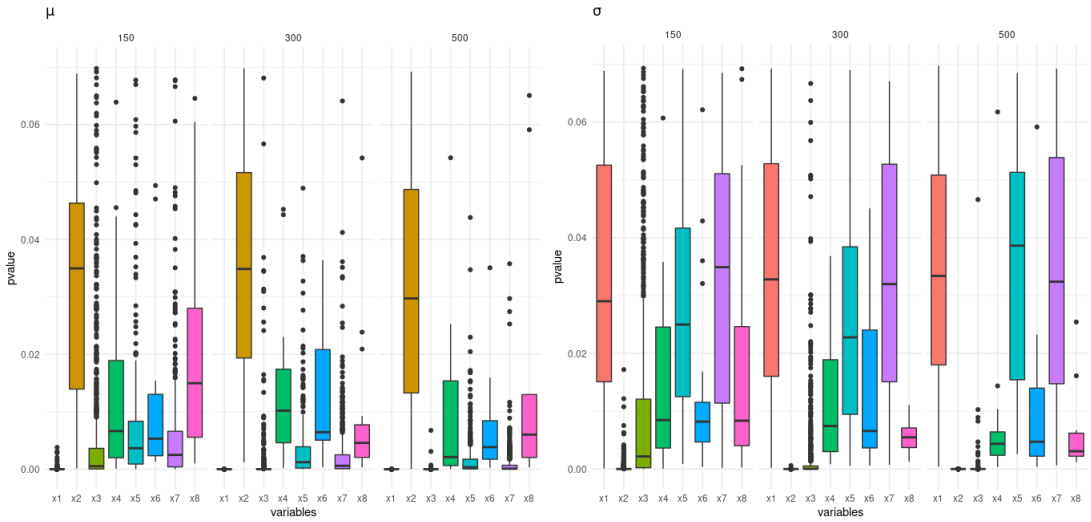
que para a Estratégia B, para $\kappa = [(log(n) + 2)/2]$ apresenta bem mais valores discrepantes se comparados com a Estratégia A.

Figura 6 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo normal com estrutura linear, com $\kappa = [(log(n) + 2)/2]$.



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 7 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo normal com estrutura linear, $\kappa = [(log(n) + 2)/2]$.



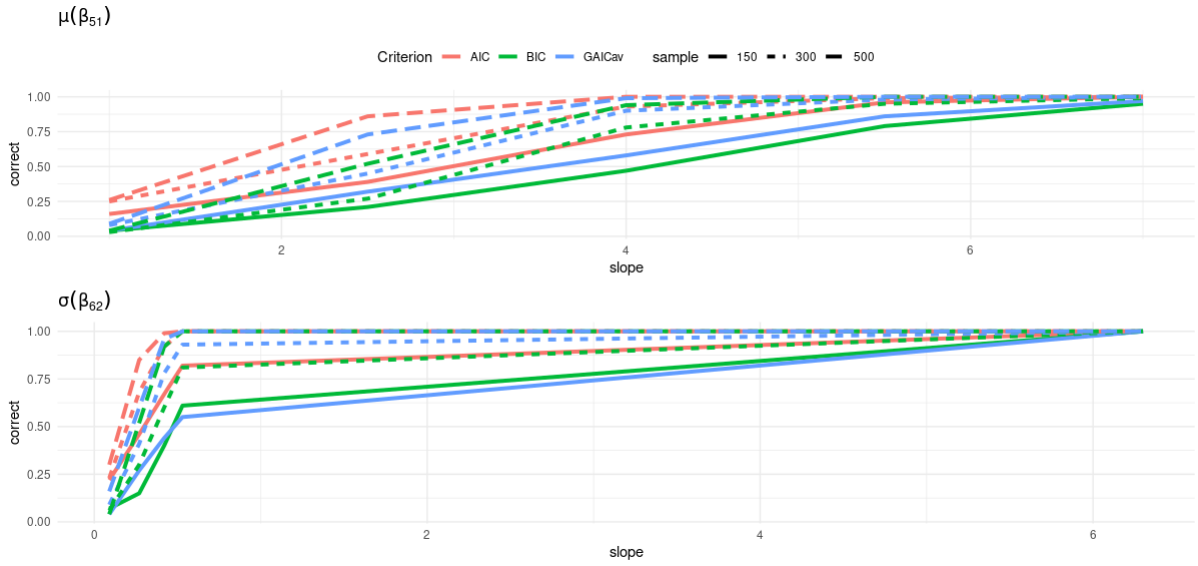
Fonte: Adaptado de (RAMIRES et al., 2021)

Com os resultados apresentados foi possível concluir que para os dois procedimentos de seleção de variáveis, os p-valores das variáveis são menores que o nível de significância, ou seja, as variáveis selecionadas são bem significativas.

Aqui, os Gráficos 8 e 9 analisam o desempenho de cada uma das estratégias comparando a inclinação do modelo e o desempenho do critério utilizado. Os resultados mostram que quando

o valor da inclinação do modelo crescer, isso vai implicar diretamente em uma maior taxa de acerto maior. Aqui foi possível perceber que para a estratégia A o critério AIC tem maiores taxas de acerto se comparados com as outras penalidades.

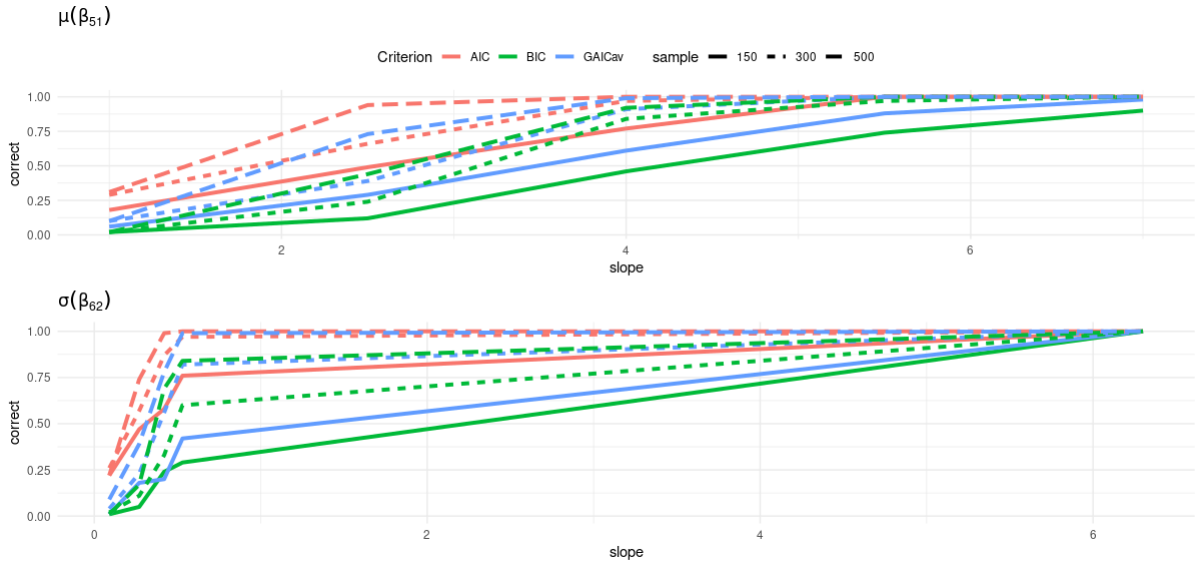
Figura 8 – Taxa de acertos da seleção de variáveis usando a Estratégia A com modelo normal, diferentes inclinações e diferentes critérios.



Fonte: Adaptado de (RAMIRES et al., 2021)

O mesmo pode-se observar para a Estratégia B, contudo, observando os Gráficos 8 e 9, percebe-se que na Estratégia A o crescimento da taxa de acerto é gradual e mais lento, se comparados com a estratégia B onde se percebe que para uma inclinação pequena, as taxas de acerto são bem maiores se comparados com a Estratégia B.

Figura 9 – Taxa de acertos da seleção de variáveis usando a Estratégia B com modelo normal, diferentes inclinações e diferentes critérios.



Fonte: Adaptado de (RAMIRES et al., 2021)

4.1.2 Resultados do estudo de simulação para dados de contagem, usando Estratégia A e a Estratégia B.

Para este cenário, foi considerado a variável aleatória $Y \sim ZIP(\mu, \sigma)$ e adotamos a mesma estrutura de regressão descrita por Ramires et al. (2021) que é dada por

$$\mu = exp[\beta_{01} + \beta_{11}x_1 + \beta_{31}x_3 + \beta_{51}x_5 + \beta_{71}x_7]$$
$$\sigma = logit[\beta_{02} + \beta_{22}x_2 + \beta_{32}x_3 + \beta_{62}x_6 + \beta_{72}x_7]$$

Com os seguintes valores verdadeiros dos parâmetros nos processos de geração de dados:

$$\mu = exp[0,4 + 0,4x_1 - 0,04x_3 + 0,04x_5 + 0,05x_7$$
$$\sigma = exp[-2,11 + 0,75x_2 - 1,85x_3 + 0,50x_6 - 0,63x_7]$$

Na Tabela 3 apresenta os percentuais de variáveis selecionadas corretamente ou não para os modelos para μ e σ, usando como procedimento para selecionar as variáveis explicativas a Estratégia A. De acordo com os resultados apresentados, para a média μ, em todos os cenários ($\kappa = 2$ (AIC), $\kappa = log(n)$ (AIC) e $\kappa = [(log(n) + 2)/2]$), a variável x_1 obteve um alto percentual de vezes que foi selecionadas corretamente dentre as 1000 replicas de Monte Carlo. Porém, para as outras variáveis autênticas, ou seja, que estão contidas no modelo, obtiveram

resultados ruins, ou seja, baixos percentuais de vezes que foi selecionada corretamente no universo de 1000 réplicas de Monte Carlo.

Esses baixos índices contrariam o que era esperado visto que as variáveis citadas fazem parte do modelo. Ramires et al. (2021) justifica que isso não significa que haja um problema no modelo considerado, mas na verdade isso pode ser explicado pelo baixo valor do coeficiente associado a $x_3(0,04)$ em relação ao coeficiente associado a $x_1(0,4)$ e, conforme explicado no cenário anterior (dados normais), o efeito das covariáveis contínuas (x_5 e x_7) uma vez que estão no mesmo intervalo das variáveis discretas autênticas.

Em relação as variáveis explicativas selecionadas para o parâmetro σ, todos os cenários apontam comportamento semelhante ao verificado no caso de dados normais, ou seja, as variáveis contínuas (x_6 e x_7) foram mal selecionadas, com índice muito baixo (menos de 10%) quando $\kappa = log(n)$ (AIC) e $\kappa = [(log(n) + 2)/2]$ e índice com uma leve tendencia de alta (entre 10% e $20\%0$).

Tabela 3 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura linear e diferentes valores para κ.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Para $\kappa = 2$, (Critério AIC)									
150	μ	68,70	14,80	11,50	12,50	14,00	14,20	12,10	11,40
300	μ	92,00	12,50	8,00	11,40	13,10	13,00	11,80	13,00
500	μ	99,00	11,10	10,30	12,60	12,30	12,30	10,70	12,30
150	σ	16,80	55,90	99,30	15,50	16,40	21,90	23,30	17,00
300	σ	15,10	80,20	100,00	15,00	15,90	26,60	32,70	15,60
500	σ	14,60	93,60	100,00	16,50	17,10	36,70	46,60	15,30
Para $\kappa = log(n)$, (Critério BIC)									
150	μ	38,70	5,00	2,40	1,80	1,60	2,50	2,90	2,00
300	μ	70,80	4,30	1,50	2,30	1,70	1,50	1,40	0,80
500	μ	92,20	2,80	0,50	0,60	1,60	2,30	0,90	0,70
150	σ	3,00	28,50	96,30	2,60	2,60	4,60	7,10	1,90
300	σ	2,20	46,40	100,00	1,50	1,60	5,20	10,00	1,60
500	σ	0,80	72,50	100,00	1,10	1,70	9,40	14,20	1,90
Para $\kappa = [(log(n) + 2)/2]$									
n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
150	μ	54,00	7,20	4,60	5,70	5,90	6,10	5,40	4,50
300	μ	84,40	5,80	2,50	4,30	4,20	5,60	4,30	4,10
500	μ	97,60	4,50	1,20	2,00	3,80	5,00	3,20	3,50
150	σ	7,40	36,10	97,70	5,10	6,60	10,80	13,90	6,80
300	σ	5,70	60,60	100,00	4,20	5,50	14,40	17,40	4,30
500	σ	4,50	82,60	100,00	4,30	4,60	19,70	24,90	3,60

Fonte: Autoria própria (2023)

Já a Tabela 4 mostram os percentuais de variáveis selecionadas corretamente para os modelos para μ e σ, usando como procedimento a Estratégia B para a seleção das covariáveis. Para o parâmetro μ e consequentemente para o parâmetro σ dado que a Estratégia B seleciona as mesmas variáveis para todos os parâmetros, em todos os cenários ($\kappa = 2$ (AIC), $\kappa = log(n)$ (BIC) e $\kappa = [(log(n) + 2)/2]$), o percentual de variáveis selecionadas corretamente é muito baixo para as variáveis x_5 e x_7, com uma leve melhora nos indices quando $\kappa = 2$ (AIC), mesmo

assim muito ruim. Já as variáveis x_1 e x_3 apresentam um auto percentual de acerto, com exceção da variável x_1 para $\kappa = log(n)$ (BIC) que apresenta uma leve queda nos percentuais de acerto.

Para o parâmetro σ, as variáveis contínuas (x_6 e x_7) foram mal selecionadas usando a estratégia B, com percentuais menores que 10% quando usado o critério BIC, já as variável x_3 obteve um excelente percentual, muito próximo de 100%.

De acordo com os percentuais de acertos para as variáveis x_5 e x_7, percebeu-se que os valores obtidos foram muito baixo, menos de 10% para os critérios AIC e BIC. Para o parâmetro σ, em todos os cenários a variável x_3 obteve percentuais de acerto muito alto (próximo a 100%). Já a variável x_2 obteve os piores resultados se comparado com a variável x_3, porém para o AIC foi bem razoável.

Aqui pela própria função da Estratégia B, se configura o mesmo problema mencionado na Seção 4.1.1 neste caso a Estratégia B selecionando variáveis de ruído.

Tabela 4 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura linear e diferentes valores para κ.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	67,40	50,00	99,80	13,80	15,00	22,80	26,60	15,10
300	μ	93,60	74,80	100,00	14,30	14,10	23,70	36,60	14,90
500	μ	98,30	89,40	100,00	13,20	14,90	31,40	44,80	14,90
150	σ	67,40	50,00	99,80	13,80	15,00	22,80	26,60	15,10
300	σ	93,60	74,80	100,00	14,30	14,10	23,70	36,60	14,90
500	σ	98,30	89,40	100,00	13,20	14,90	31,40	44,80	14,90
				Para $\kappa = log(n)$, (Critério BIC)					
150	μ	22,70	10,80	88,40	0,90	1,10	2,00	3,40	0,70
300	μ	44,30	23,90	99,90	0,40	0,50	2,00	2,20	0,20
500	μ	72,30	37,20	100,00	0,30	0,30	2,30	3,10	0,10
150	σ	22,70	10,80	88,40	0,90	1,10	2,00	3,40	0,70
300	σ	44,30	23,90	99,90	0,40	0,50	2,00	2,20	0,20
500	σ	72,30	37,20	100,00	0,30	0,30	2,30	3,10	0,10
				Para $\kappa = [(log(n) + 2)/2]$					
150	μ	40,00	23,20	96,90	3,20	3,20	5,80	7,20	4,30
300	μ	70,40	44,40	99,90	2,10	2,80	7,10	9,60	3,30
500	μ	90,70	67,80	100,00	1,70	2,10	8,20	12,60	1,80
150	σ	40,00	23.20	96,90	3,20	3,20	5,80	7,20	4,30
300	σ	70,40	44,40	99,90	2,10	2,80	7,10	9,60	3,30
500	σ	90,70	67,80	100,00	1,70	2,10	8,20	12,60	1,80

Fonte: Autoria própria (2023)

Quando comparou-se os procedimentos Estratégia A e Estratégia B para os parâmetros μ e σ, foi possível concluir que a Estratégia A obteve melhores percentuais na maioria dos cenários considerados para a variável x_1 e x_3 para o parâmetro μ e x_2 para o parâmetro σ.

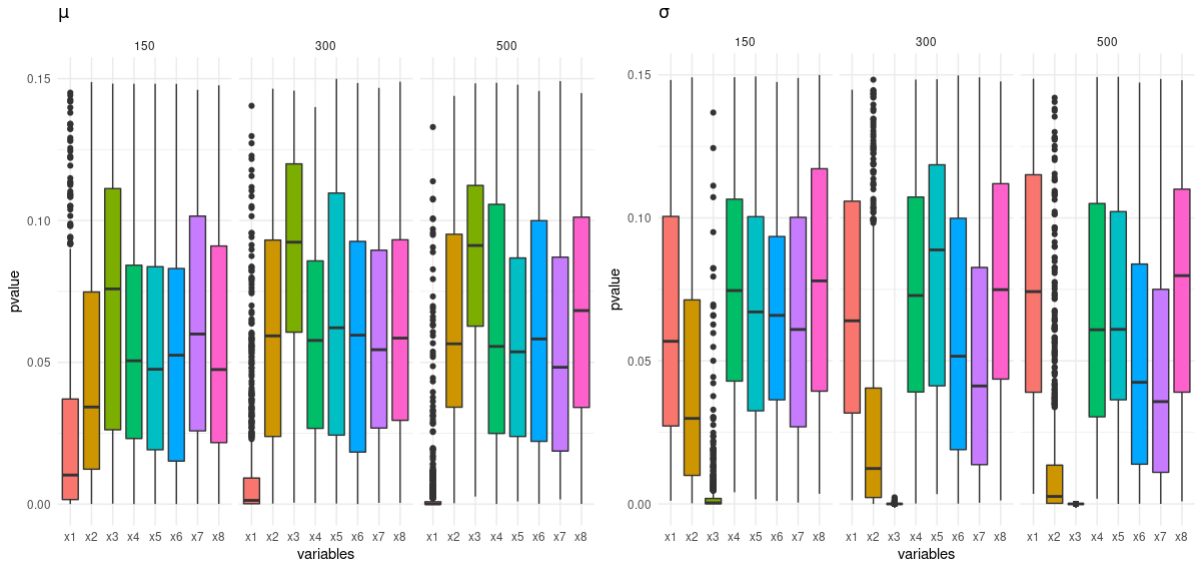
E na comparação das variáveis x_3 e x_7, que afetam diretamente os dois modelos (μ e σ), foi possível concluir que x_3 obteve excelentes percentuais para os procedimentos de seleção Estratégia A e Estratégia B.

Os Boxplot 10, 11 e 12 apresenta os p-valores calculados no estudo de simulação, usando

a Estratégia A como procedimento de seleção de variáveis explicativas e de acordo com os resultados apresentados, o critério AIC obteve uma maior variabilidade nos p-valores e o critério BIC os menores p-valores e uma variação muito pequena se comparada com o critério AIC. Além disso é possível perceber que os p-valores das variáveis que estão contidas no modelo são bem menores, o que de fato era esperado.

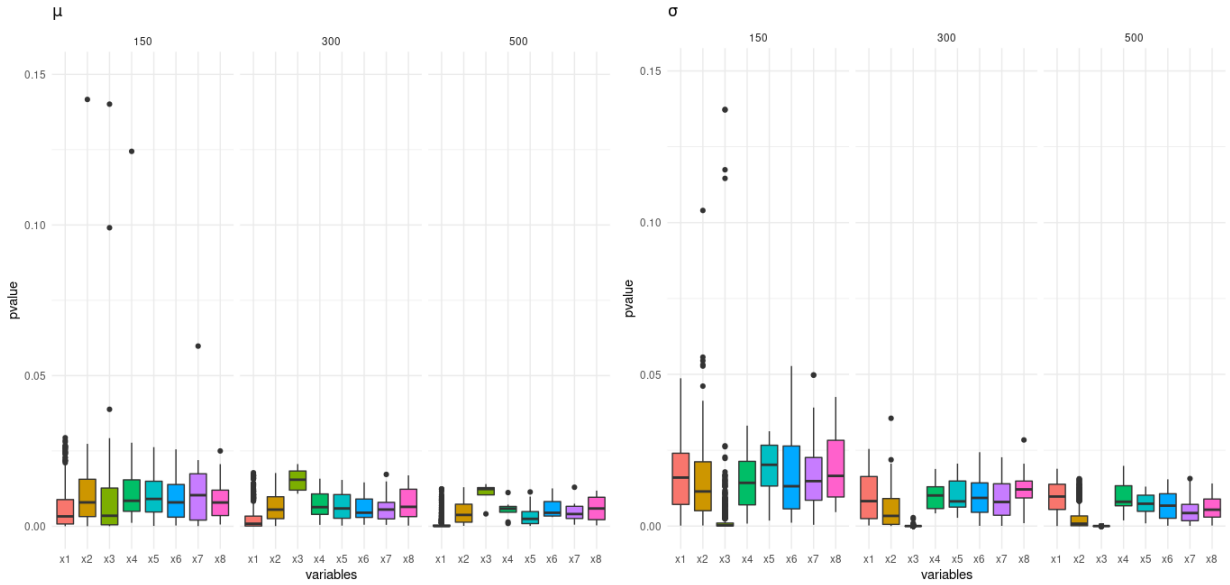
Um ponto importante é para o critério BIC, que apresenta valores muito baixo para os p-valores, o que implica que as variáveis são significativas para o modelo ajustado.

Figura 10 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura linear, com penalidade $\kappa = 2$(AIC).



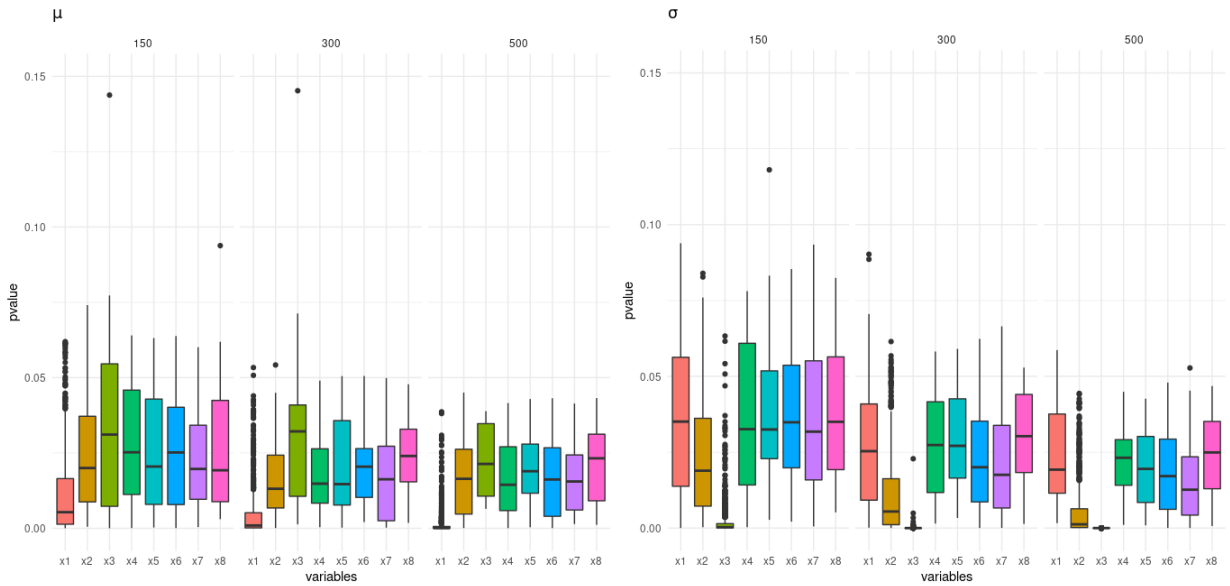
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 11 – P-valores da seleç ão de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura linear, com penalidade $\kappa = \log(n)$(BIC).



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 12 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura linear, com penalidade $\kappa = [(log(n) + 2)/2]$.

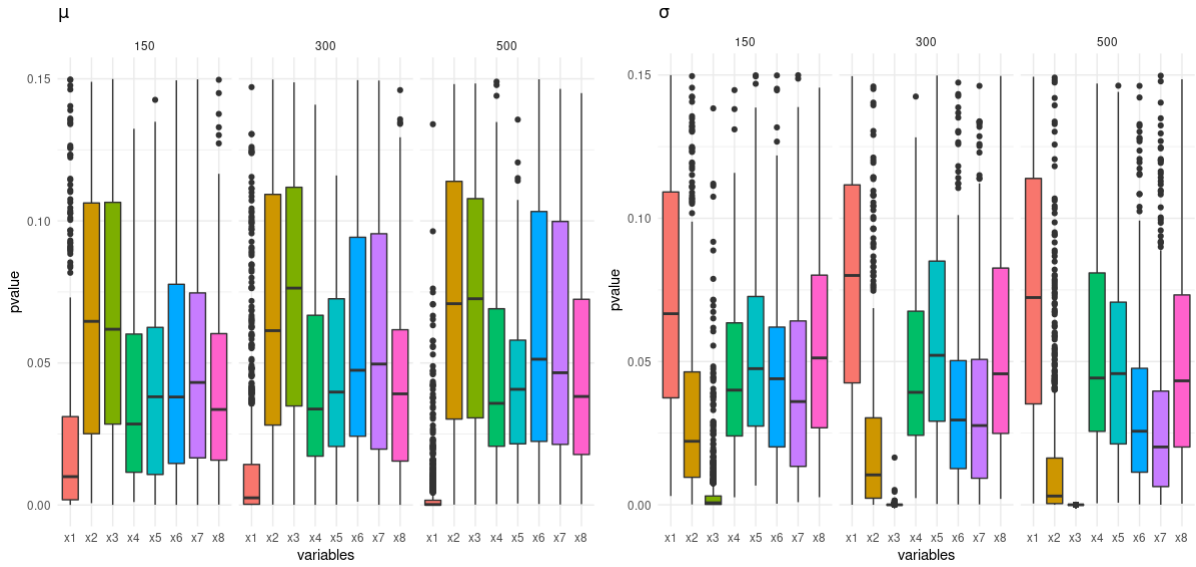


Fonte: Adaptado de (RAMIRES et al., 2021)

De acordo com os Boxplot 13, 14 e 15 que mostra os p-valores encontrados para cada variável no estudo de simulação usando a Estratégia B como procedimento de seleção de variáveis. Os resultados mostram que para as variáveis que estão no modelo, os p-valores são bem menores, principalmente a variável x_1 para o submodelo μ e as variáveis x_2 e x_3

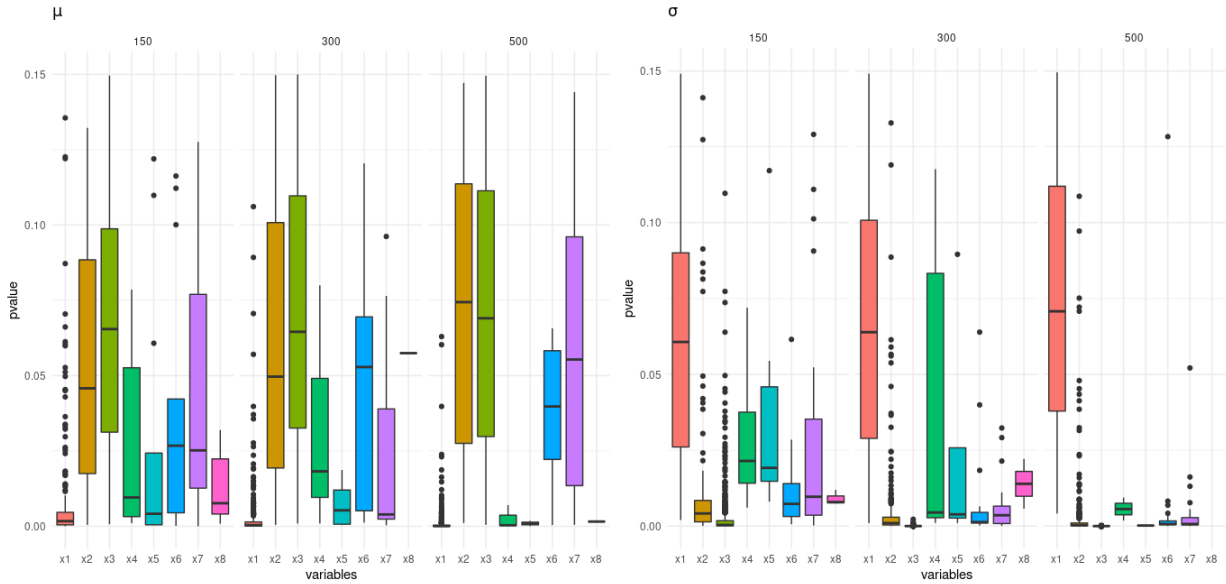
para o submodelo σ. Por outro lado, para as variáveis que não estão contidas nos modelos, a variabilidade é muito grande nos três critérios utilizados.

Figura 13 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura linear, com penalidade $\kappa = 2$(AIC).



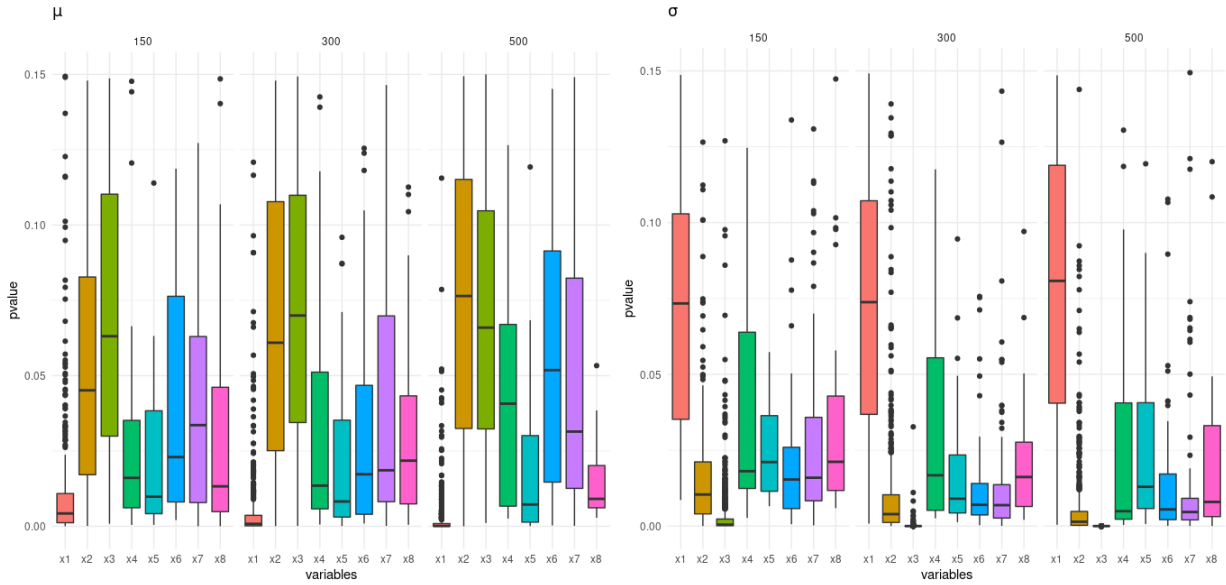
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 14 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura linear, com penalidade $\kappa = \log(n)$(BIC).



Fonte: Adaptado de (RAMIRES et al., 2021)

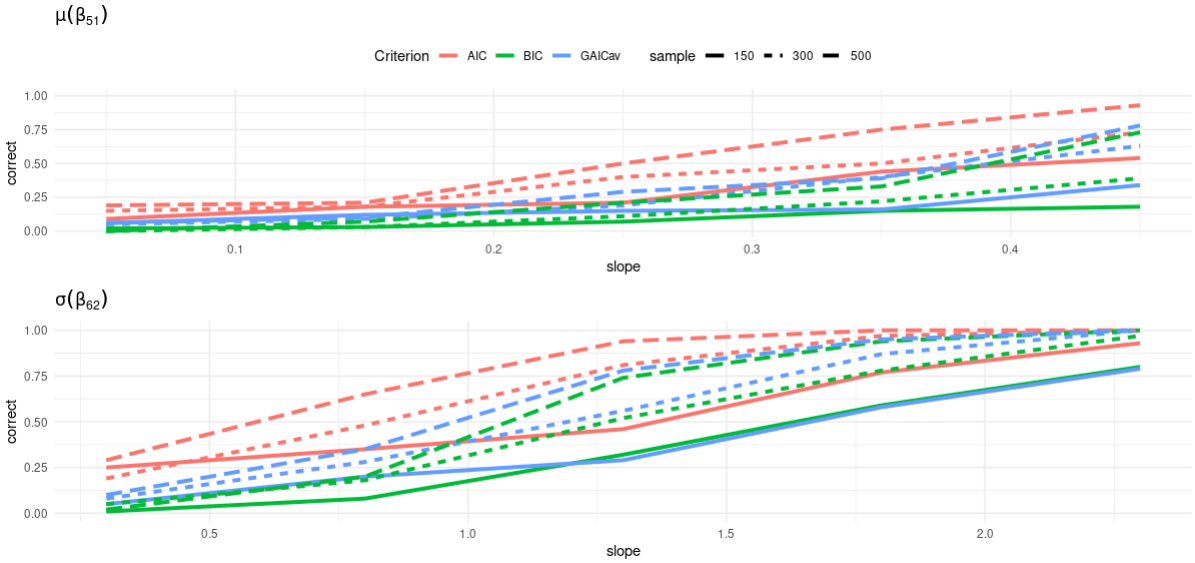
Figura 15 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura linear, com penalidade $\kappa = [(log(n) + 2)/2]$.



Fonte: Adaptado de (RAMIRES et al., 2021)

O Gráfico 16 mostra os resultados da performance dos critérios AIC, BIC e GAIC, quando os valores da inclinação do modelo é aumentado. Para o procedimento de seleção Estratégia A, concluiu-se que o desempenho dos critérios de seleção melhoram significativamente quando o valor da inclinação do modelo aumenta.

Figura 16 – Taxa de acertos da seleção de variáveis usando a Estratégia A com modelo normal, diferentes inclinações e diferentes penalidades.

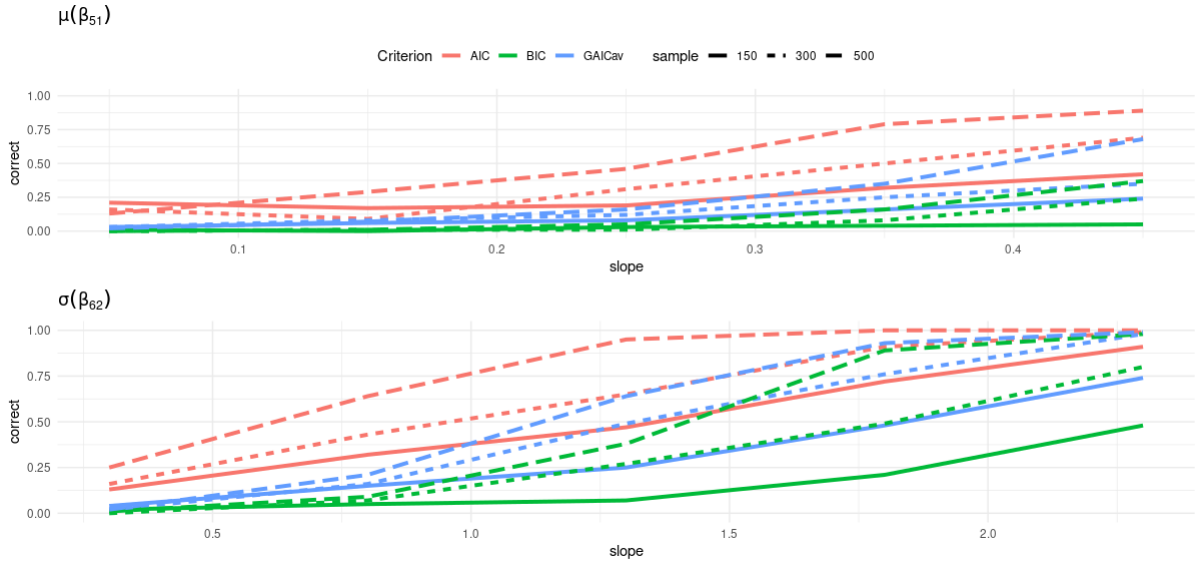


Fonte: Adaptado de (RAMIRES et al., 2021)

O Gráfico 17 apresenta os resultados para a Estratégia B do desempenho dos critérios

de seleção quando a inclinação do modelo é modificada. Os resultados são semelhantes aos resultados apresentados no Gráfico 16 e portanto a análise é a mesma.

Figura 17 – Taxa de acertos da seleção de variáveis usando a Estratégia B com modelo normal, diferentes inclinações e diferentes penalidades.



Fonte: Adaptado de (RAMIRES et al., 2021)

4.2 SELEÇÃO DE VARIÁVEIS EXPLICATIVAS - ESTRATÉGIA A E ESTRATÉGIA B, EM MODELOS COM ESTRUTURA NÃO LINEAR

Neste cenário, foi considerado modelos de regressão com estrutura linear e também modelos de regressão com estruturas não lineares. A composição de modelos não lineares deve ser caracterizados por funções $f(X, \theta)$ não lineares em pelo menos um dos parâmetros θ_i, ou seja, para que um modelo seja não linear uma condição obrigatória é que ao menos uma das derivadas da $f(x, \theta)$ com relação a θ_i dependa de ao menos um dos parâmetros θ_i.

Para este cenário, foi adotado o mesmo esquema proposto por Ramires et al. (2021), a saber: i) as variáveis x_1, x_5 e x_7 só serão consideradas no processo de geração para μ e as variáveis x_2 e x_6 so serão consideradas no processo de geração para σ; ii) x_7 será gerada de tal forma que seu efeito em μ tenha uma forma crescente, decrescente, crescente.

4.2.1 Resultados dos estudos de simulação dos procedimentos de seleção de variáveis explicativas - Estratégia A e Estratégia B, para modelo de regressão normal e com uma estrutura não linear

Considerou-se a variável aleatória $Y \sim N(\mu, \sigma)$ e para a variável explicativa x_7 foi colocado um efeito não linear para o parâmetro μ. Foi adotado a estrutura de regressão descrita por Ramires et al. (2021) dada por

$$\mu = \beta_{01} + \beta_{11}x_1 + \beta_{31}x_3 + \beta_{51}x_5 + s(x_7)$$
$$\sigma = exp[\beta_{02} + \beta_{22}x_2 + \beta_{32}x_3 + \beta_{62}x_6]$$

em que $s(.)$ é uma p-spline usada para modelar a relação entre a variável explicativa x_7 e o parâmetro μ (média).

Nos processos de geração de dados, foi considerado como valores verdadeiros dos parâmetros nos processos os seguintes:

$$\mu = 40 + 5x_1 - 4x_3 + 2,5x_5 - 10.sen.(0,2x_7\pi)$$
$$\sigma = exp[1,6 + 0,6x_2 - 0,35x_3 + 0,03x_6]$$

Os resultados dos estudos para este cenário são apresentados na Tabela 5, referentes a Estratégia A e os percentuais de seleção correta para variáveis explicativas dos modelos para os parâmetros μ e σ são mostrados, usando os três tipos de penalidades consideradas nesta pesquisa (ou seja, três valores diferentes para κ).

De acordo com os resultados mostrados, podemos perceber que para a Estratégia A, algumas variáveis (x_1 e x_3 para μ e x_6 para σ) obtiveram resultados similares aos apresentados na Seção 4.1.1 que trata do estudo de simulação em modelos normais sem o uso de funções suavizadoras, nesse caso, a análise deve ser a mesma. No entanto, para a variável x_2, x_5 e x_7, ou seja, onde se esperava um bom índice de seleção correta, obtivemos índices muito baixo e para as variáveis de ruido, se esperaria baixo percentual de seleção correta, os resultados mostraram o contrário. Uma justificativa para os baixos percentuais obtidos pela variáveis x_5 e x_7 pode ser os baixos valores para os seus coeficientes.

Tabela 5 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo Normal com estrutura não linear e diferentes valores para κ.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	97,80	100,00	12,50	99,20	10,90	43,10	9,80	9,40
300	μ	99,70	100,00	10,90	100,00	8,90	68,30	8,90	9,20
500	μ	100,00	100,00	11,90	100,00	7,60	84,90	9,50	9,10
150	σ	26,50	17,50	99,90	90,40	17,20	15,90	18,60	17,30
150	σ	24,00	15,70	100,00	99,40	17,30	15,80	16,20	16,60
150	σ	24,90	16,40	100,00	100,00	14,00	15,80	17,20	15,70
				Para $\kappa = log(n)$, (Critério BIC)					
300	μ	79,80	98,90	1,30	93,50	1,30	14,00	2,10	2,10
300	μ	97,50	100,00	1,20	99,90	1,50	31,10	0,40	0,40
300	μ	100,00	100,00	1,00	100,00	0,70	50,50	0,60	0,70
150	σ	3,90	3,20	99,00	69,80	4,20	2,70	3,30	2,80
300	σ	1,40	1,80	100,00	96,60	2,40	2,60	3,50	2,10
500	σ	0,70	1,10	100,00	99,60	1,10	1,30	2,30	1,90
				Para $\kappa = [(log(n) + 2)/2]$					
150	μ	91,10	99,80	5,10	96,50	3,90	26,30	3,60	3,90
300	μ	99,20	100,00	4,20	100,00	2,90	48,20	2,20	2,10
500	μ	100,00	100,00	2,70	100,00	2,40	66,70	1,80	2,20
150	σ	7,70	6,70	99,70	81,90	6,60	6,10	8,40	6,80
300	σ	5,50	6,00	100,00	98,10	4,20	4,90	6,50	5,40
500	σ	3,80	3,70	100,00	100,00	5,00	4,90	5,00	5,50

Fonte: Autoria própria (2023)

Na Tabela 6, referentes a Estratégia B, estão presentes os percentuais de seleção correta para variáveis explicativas dos modelos par os parâmetros μ e σ, usando a Estratégia B e os três tipos de penalidades consideradas nesta pesquisa. Os resultados apresentados para a Estratégia B mostrado na **Tabela** 6 são similares ao modelo com estrutura linear apresentado na Seção 4.1.1, o implica portanto as mesmas conclusões e o mesmo problema citado.

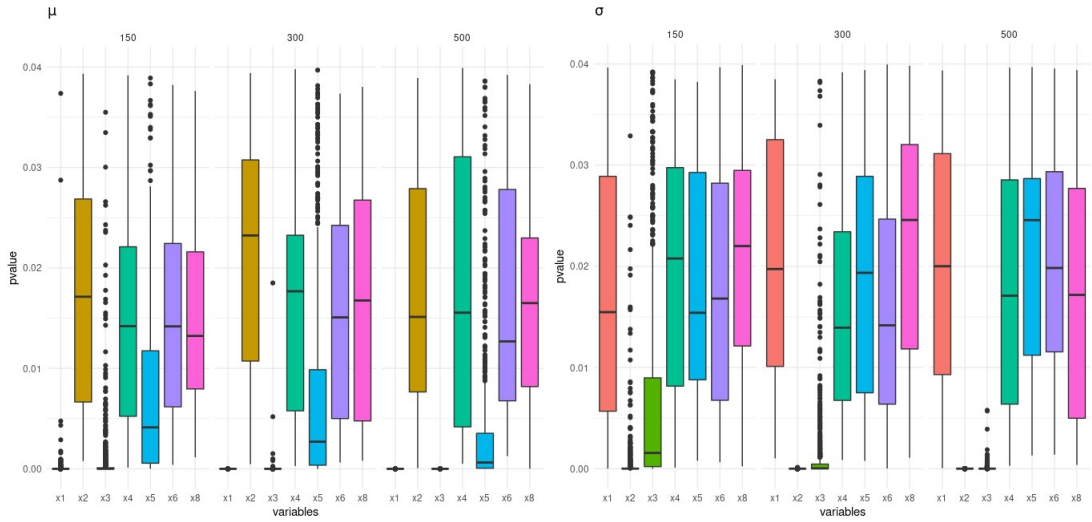
Tabela 6 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo Normal com estrutura não linear e diferentes valores para κ.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	100,00	99,90	99,90	17,50	54,60	17,50	98,80	17,90
300	μ	100,00	100,00	100,00	13,30	74,10	16,30	100,00	16,70
500	μ	100,00	100,00	100,00	14.70	91.30	15.20	100,00	14,40
150	σ	100,00	99,90	99,90	17,50	54,60	17,50	98,80	17,90
300	σ	100,00	100,00	100,00	13,30	74,10	16,30	100,00	16,70
500	σ	100,00	100,00	100,00	14,70	91,30	15,20	100,00	14,40
				Para $\kappa = log(n)$, (Critério BIC)					
150	μ	99,00	94,60	96,80	0,80	12,00	1,10	75,60	1,00
300	μ	100,00	100,00	100,00	0,40	23,30	0,30	98,00	0,20
500	μ	100,00	100,00	100,00	0,50	42,40	0,10	100,00	0,20
150	σ	99,00	94,60	96,80	0,80	12,00	1,10	75,60	1,00
300	σ	100,00	100,00	100,00	0,40	23,30	0,30	98,00	0,20
500	σ	100,00	100,00	100,00	0,50	42,40	0,10	100,00	0,20
				Para $\kappa = [(log(n) + 2)/2]$					
150	μ	100,00	98,30	99,70	4,70	27,00	4,80	92,60	4,20
300	μ	100,00	100,00	10,00	2,80	43,90	2,80	99,50	2,40
500	μ	100,00	100,00	100,00	2,00	66,40	2,40	100,00	2,00
150	σ	100,00	98,30	99,70	4,70	27,00	4,80	92,60	4,20
300	σ	100,00	100,00	100,00	2,80	43,90	2,80	99,50	2,40
500	σ	100,00	100,00	100,00	2,00	66,40	2,40	100,00	2,00

Fonte: Autoria própria (2023)

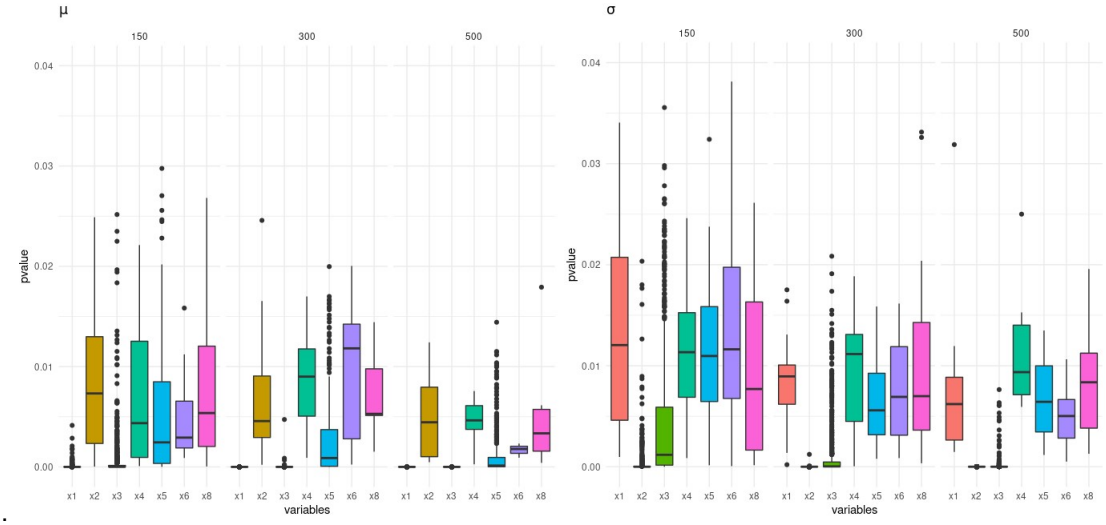
Observando as Figuras 18, 19 e 20, onde foi utilizado a Estratégia A para selecionar variáveis, percebe-se que para as variáveis que fazem parte dos submodelos para μ e σ os p-valores são muito baixos, próximos de zero, com exceção da variável x_3 para o submodelo para μ e a variável x_5 para o submodelo para σ, que embora significativo para o modelo, obtiveram p-valores maiores, enquanto que para as demais variáveis explicativas os p-valores são bem maiores e apresentam grande variabilidade.

Figura 18 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo Normal com estrutura não linear e com penalidade $\kappa = 2$(AIC).



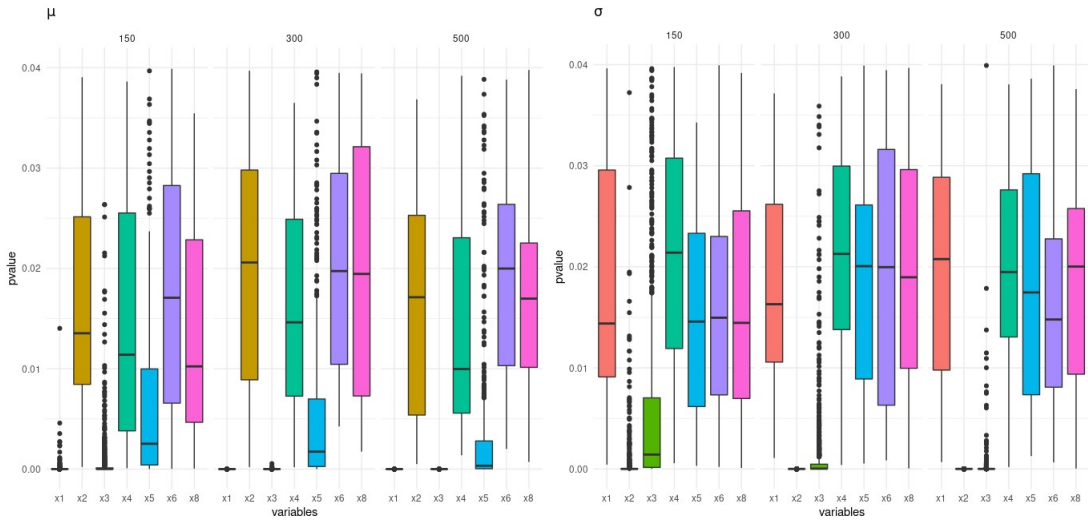
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 19 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo Normal com estrutura não linear e com penalidade $\kappa = log(n)$(BIC)



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 20 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo Normal com estrutura não linear e com penalidade $\kappa = [(log(n) + 2)/2]$.

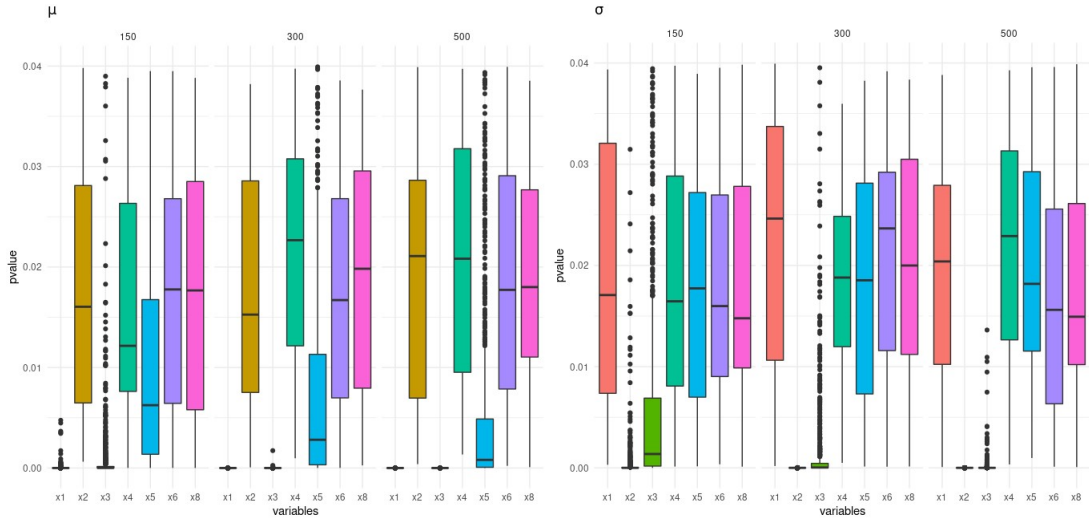


Fonte: Adaptado de (RAMIRES et al., 2021)

Para as três penalizações ($\kappa = 2$ (AIC), $\kappa = log(n)$ (BIC) e $\kappa = [(log(n) + 2)/2]$)) utilizadas, percebe-se que o critério BIC apresenta p-valoees menores e com menor variabilidade.

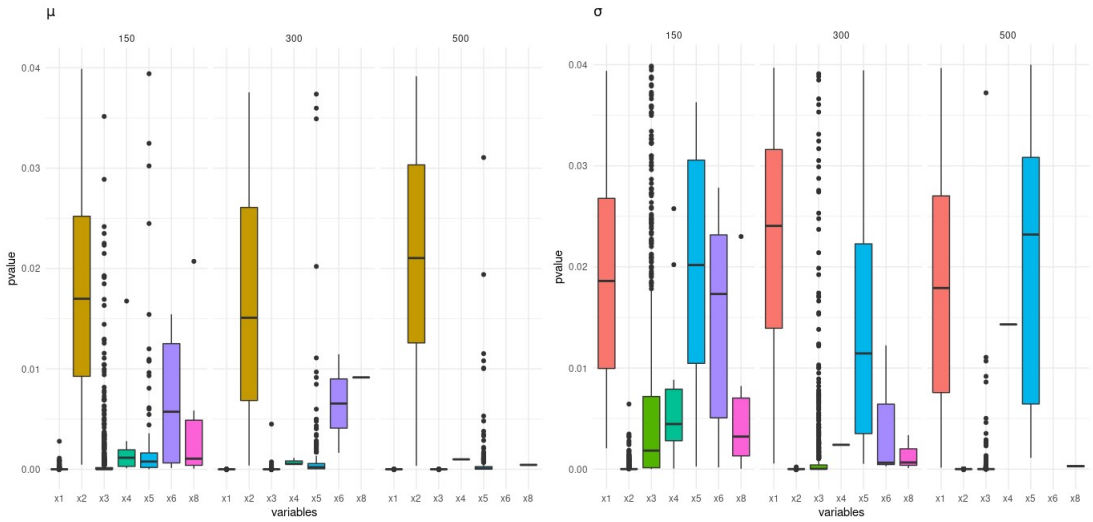
As Figuras 21, 22 e 23 mostram resultados similares quando usado a Estratégia B, se comparado com o mesmo cenário usando a Estratégia A para selecionar variáveis explicativas, ou seja, quando observamos os p-valores obtidos, percebemos que os resultados são similares aos p-valores obtidos quando usado a Estratégia B, neste mesmo cenário, com exceção do critério BIC que apresentou resultados maiores que os apresentados na Estratégia A.

Figura 21 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo Normal com estrutura não linear e com penalidade $\kappa = 2$(AIC).



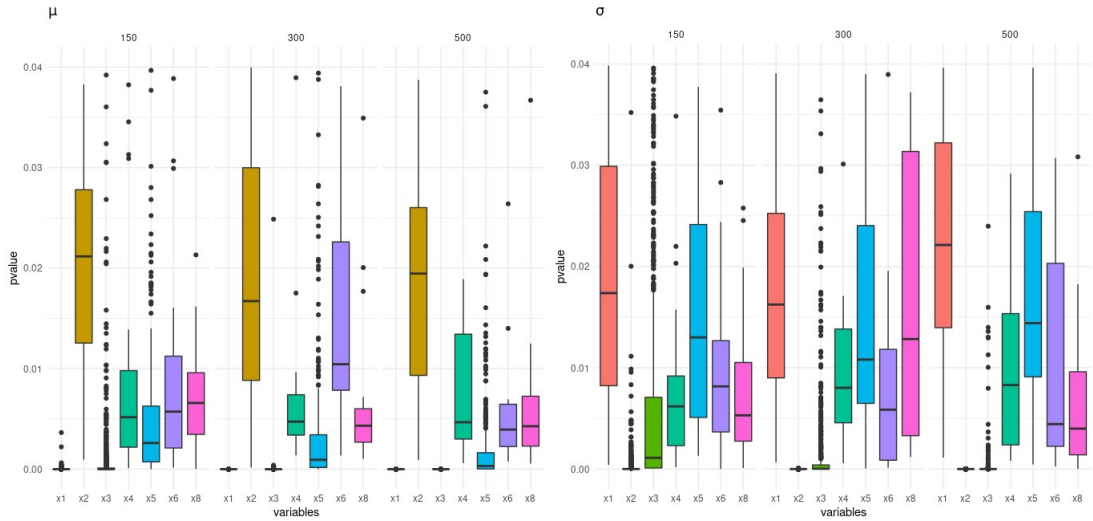
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 22 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo Normal com estrutura não linear e com penalidade $\kappa = log(n)$(BIC).



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 23 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo Normal com estrutura não linear e com penalidade $\kappa = [(log(n) + 2)/2]$.



Fonte: Adaptado de (RAMIRES et al., 2021)

4.2.2 Resultados dos estudos de simulação dos procedimentos de seleção de variáveis explicativas - Estratégia A e Estratégia B, para modelo de regressão ZIP e com uma estrutura não linear

Aqui consideramos a variável aleatória $Y \sim ZIP(\mu, \sigma)$ e a variável explicativa x_7 com um efeito não linear para o parâmetro μ. Seja estrutura de regressão descrita por Ramires et al.

(2021) dada por

$$\mu = exp[\beta_{01} + \beta_{11}x_1 + \beta_{31}x_3 + \beta_{51}x_5 + s(x_7)] \tag{4.3}$$

$$\sigma = logit[\beta_{02} + \beta_{22}x_2 + \beta_{32}x_3 + \beta_{62}x_6] \tag{4.4}$$

em que $s(.)$ é uma $p - spline$ usada para modelar a relação entre a variável explicativa x_7 e o parâmetro μ (média).

Como valores verdadeiros dos parâmetros nos processos de geração de dados temos os seguintes:

$$\mu = \exp[0,4 + 0,4x_1 + 0,04x_3 + 0,04x_5 + 0,2.\sin.(0,2\pi_7)] \tag{4.5}$$

$$\sigma = logit[-2,11 + 0,75x_2 + 1,85x_3 + 0,5x_6] \tag{4.6}$$

Neste cenário foi relizado um estudo de simulação com 1000 réplicas de Monte Carlos e três tamanhos amostrais. As Tabelas 7 e 8 apresentam os percentuais de especificação correta para cada variável explicativa referente a seleção de variáveis esplicativa utilizando a Estratégia A e a Estratégia B respectivamente. De acordo com os resultados apresentados para o modelo ZIP com estrutura não linear, percebeu-se que os resultados são similares aos resultados obtidos na simulação do modelo ZIP com estrutura linear, e neste caso, a análise e interpretação dos resultados deve seguir a mesma ideia ou seja, é análoga.

Tabela 7 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura não linear e diferentes valores para κ.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Para $\kappa = 2$, (Critério AIC)									
150	μ	$72,20$	$15,20$	$8,80$	$13,50$	$12,70$	$13,90$	$24,70$	$14,00$
300	μ	$95,90$	$12,80$	$7,40$	$12,70$	$12,00$	$12,70$	$26,80$	$13,30$
500	μ	$99,30$	$10,00$	$9,10$	$14,70$	$13,50$	$12,40$	$29,20$	$15,20$
150	σ	$18,10$	$53,40$	$98,30$	$16,20$	$18,70$	$23,20$	$23,00$	$17,40$
300	σ	$16,80$	$78,50$	$100,00$	$15,70$	$16,20$	$28,40$	$23,30$	$13,90$
500	σ	$15,70$	$93,30$	$100,00$	$13,80$	$15,40$	$37,40$	$22,60$	$14,30$
Para $\kappa = log(n)$, (Critério BIC)									
150	μ	$50,10$	$3,90$	$3,40$	$2,50$	$3,70$	$2,70$	$3,60$	$2,00$
300	μ	$80,00$	$2,60$	$0,40$	$1,00$	1.40	$1,50$	$2,70$	$2,00$
500	μ	$96,00$	$2,60$	$0,20$	$1,10$	$0,80$	$2,00$	$2,20$	$0,90$
150	σ	$2,70$	$25,10$	$94,90$	$2,40$	$2,00$	$4,60$	$3,80$	$3,10$
300	σ	$2,80$	$48,30$	$99,80$	$1,40$	$2,30$	$7,00$	$2,50$	$1,80$
500	σ	$1,40$	$68,30$	$100,00$	$1,60$	$1,60$	$8,60$	$1,00$	$0,70$
Para $\kappa = [(log(n) + 2)/2]$									
150	μ	$62,00$	$8,90$	$4,20$	$5,30$	$5,50$	$6,30$	$9,80$	$6,00$
300	μ	$89,20$	$5,10$	$1,90$	$3,10$	$4,10$	$4,50$	$8,00$	$3,70$
500	μ	$98,10$	$4,10$	$1,40$	$2,40$	$2,80$	$4,30$	$11,00$	$2,70$
150	σ	$7,80$	$37,80$	$97,70$	$6,60$	$7,10$	$12,20$	$7,80$	$6,70$
300	σ	$5,60$	$63,90$	$100,00$	$5,10$	$5,20$	$14,30$	$5,20$	$4,50$
500	σ	$5,40$	$84,80$	$100,00$	$5,10$	$4,70$	$16,70$	$4,10$	$4,60$

Fonte: Autoria própria (2023)

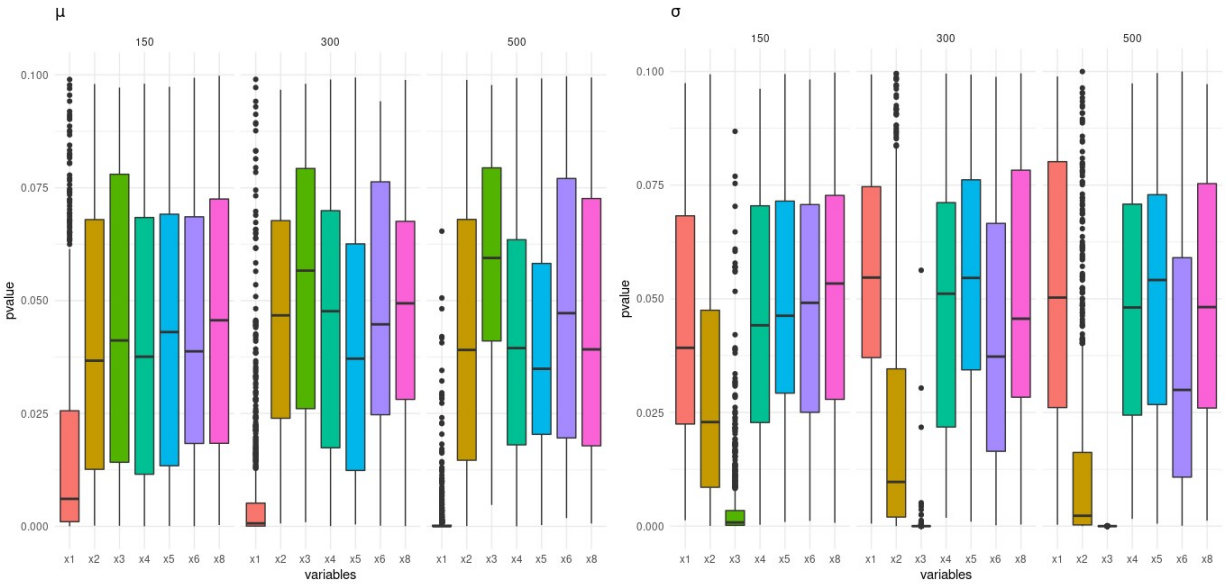
Tabela 8 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura não linear e usando $\kappa = 2$ (AIC), $\kappa = log(n)$ (BIC) e $\kappa = [(log(n) + 2)/2]$.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	71,90	49,80	99,10	14,10	17,90	22,00	27,70	18,80
300	μ	94,70	74,30	100,00	15,50	15,00	23,30	31,40	14,80
500	μ	99,40	89,90	100,00	14,00	15,10	32,50	32,30	16,60
150	σ	71,90	49,80	99,10	14,10	17,90	22,00	27,70	18,80
300	σ	94,70	74,30	100,00	15,50	15,00	23,30	31,40	14,80
500	σ	99,40	89,90	100,00	14,00	15,10	32,50	32,30	16,60
				Para $\kappa = log(n)$, (Critério BIC)					
150	μ	24,50	9,80	87,10	0,60	0,60	1,60	1,00	0,40
300	μ	55,40	20,00	99,40	0,30	0,40	1,30	1,10	0,30
500	μ	82,20	40,10	100,00	0,20	2,30	0,40	0,60	1,40
150	σ	24,50	9,80	87,10	0,60	0,60	1,60	1,00	0,40
300	σ	55,40	20,00	99,40	0,30	0,40	1,30	1,10	0,30
500	σ	82,20	40,10	100,00	0,20	2,30	0,40	0,60	0,00
				Para $\kappa = [(log(n) + 2)/2]$					
150	σ	49,30	24,40	95,40	3,50	4,70	6,80	5,80	3,90
300	σ	79,80	40,30	100,00	2,10	2,70	6,70	5,00	3,80
500	σ	95,80	65,50	10,00	2,40	1,70	8,70	4,30	1,50
150	σ	49,30	24,40	95,40	3,50	4,70	6,80	5,80	3,90
300	σ	79,80	40,30	100,00	2,10	2,70	6,70	5,00	3,80
500	σ	95,80	65,50	100,00	2,40	1,70	8,70	4,30	1,50

Fonte: Autoria própria (2023)

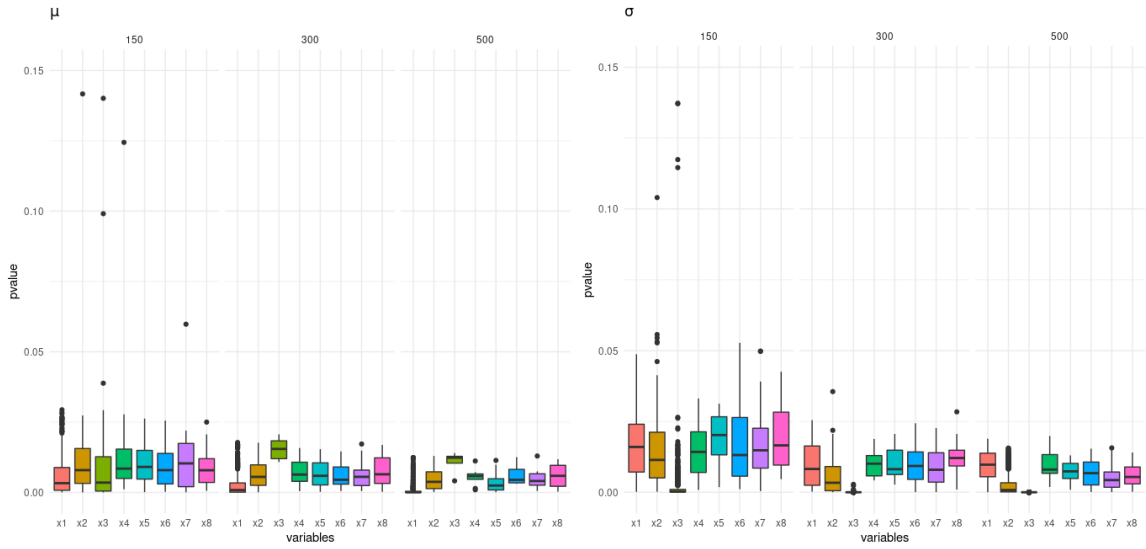
Aqui os Gráficos 24, 25 e 26 apresentam os p-valores obtidos no processo de seleção de variáveis explicativas para os três tamanhos amostrais nas 1000 réplicas de Monte Carlo, usando a Estratégia A. Podemos perceber que para o critério AIC obteve-se valores maiores e com uma maior variabilidade, enquanto que para os critérios BIC e GAIC percebe-se uma variabilidade bem menor, principalmente para o critério BIC.

Figura 24 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura não linear e com penalidade $\kappa = 2$(AIC).



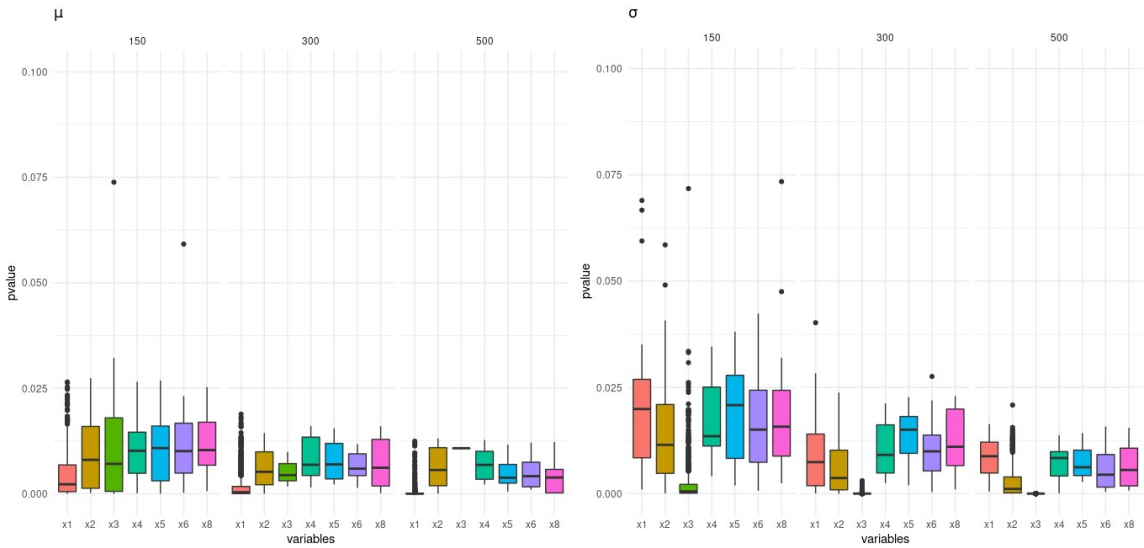
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 25 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura não linear e com penalidade $\kappa = log(n)$(BIC).



Fonte: Adaptado de (RAMIRES et al., 2021)

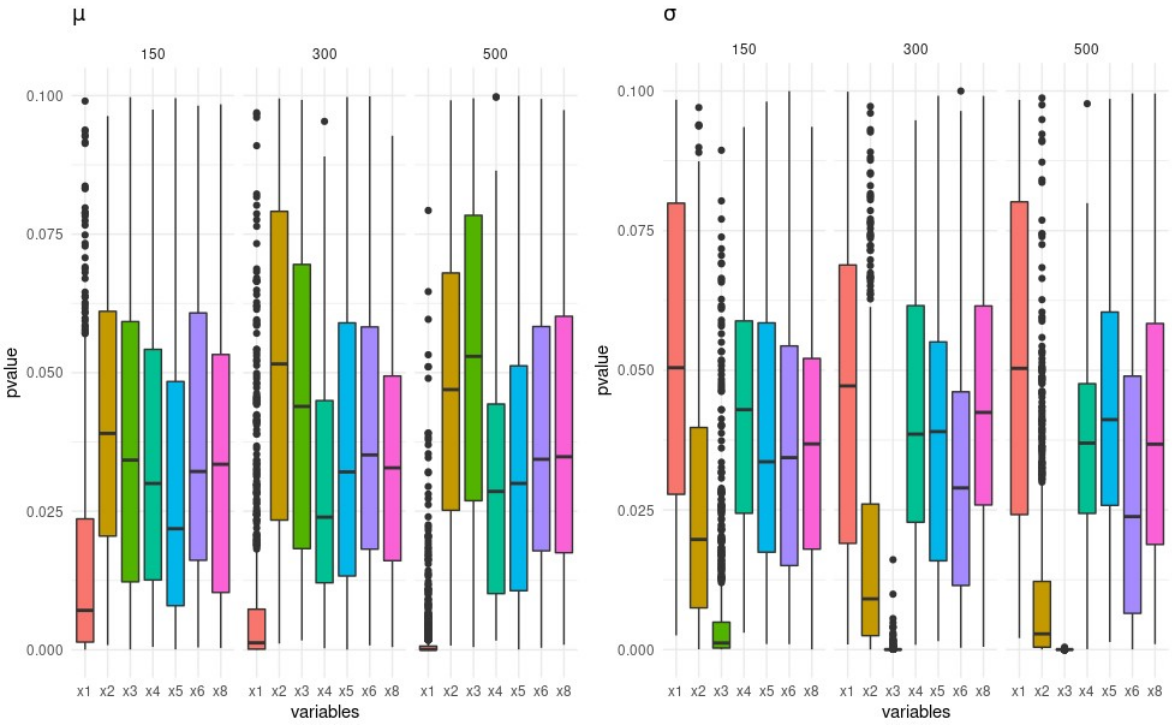
Figura 26 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia A em um modelo ZIP com estrutura não linear e com penalidade $\kappa = [(log(n) + 2)/2]$.



Fonte: Adaptado de (RAMIRES et al., 2021)

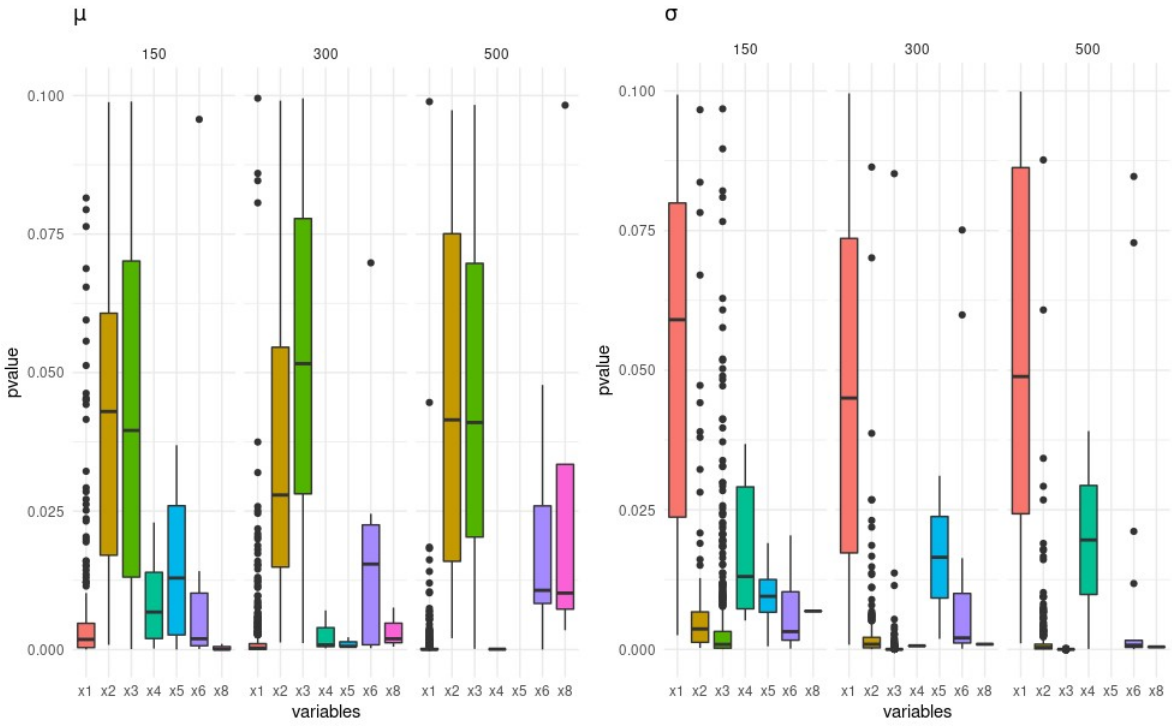
Os Boxplot 27, 28 e 29 que apresentam os p-valores obtidos no processo de seleção de variáveis explicativas para os três tamanhos amostrais nas 1000 réplicas de Monte Carlos, usando a Estratégia B. De acordo com os resultados apresentados, diferentemente da estratégia A, para os três critérios adotados foi possível perceber que para as variáveis não contidas tanto no submodelo para μ como para o submodelo σ, a variabilidade dos p-valores é bem maior. O critério AIC tem menor variabilidade.

Figura 27 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura não linear e com penalidade $\kappa = 2$ (AIC).



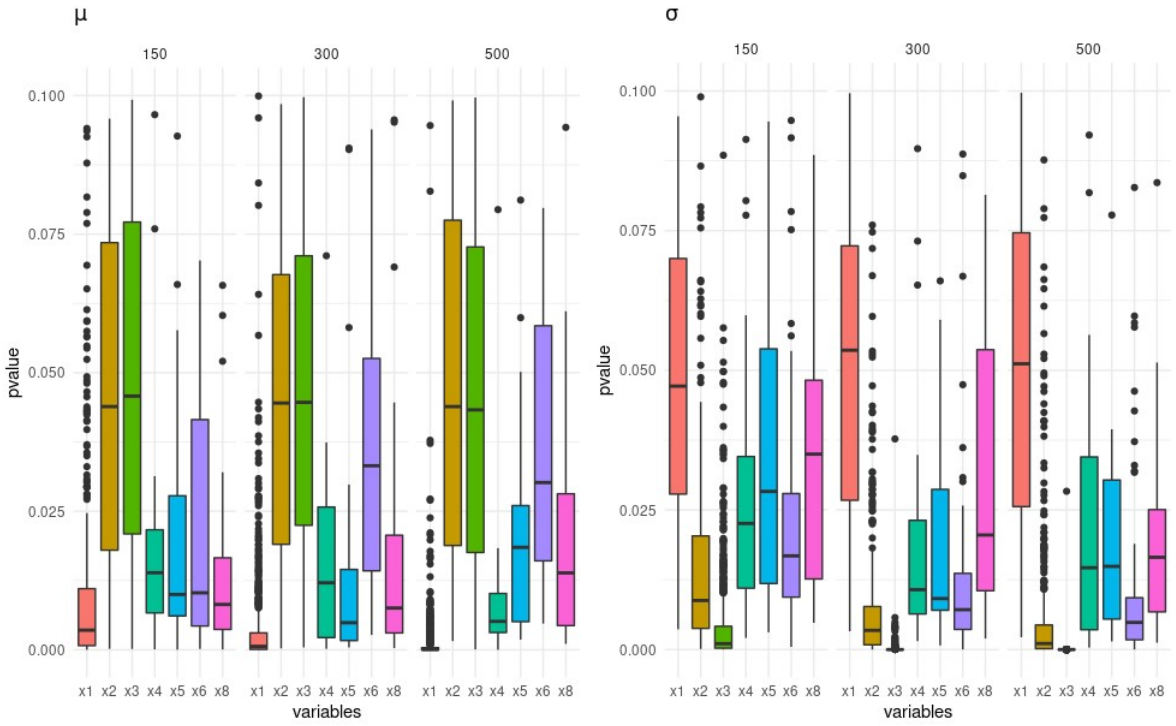
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 28 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura não linear e com penalidade $\kappa = log(n)$ (BIC).



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 29 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia B em um modelo ZIP com estrutura não linear e com penalidade $\kappa = [(log(n) + 2)/2]$.



Fonte: Adaptado de (RAMIRES et al., 2021)

4.3 SIMULAÇÃO DOS PROCEDIMENTOS DE SELEÇÃO DE VARIÁVEIS EXPLICATIVAS - ESTRATÉGIA COM ESTRUTURA LINEAR - ESTRATÉGIA C.

Neste cenário é apresentado uma proposta para a realização do estudo de simulação usando Estratégia A e a Estratégia B. Esta proposta, denominada de Estratégia C consiste em uma combinação da Estratégia B com a Estratégia A.

Os procedimentos usados para esta combinação é descrito da seguinte forma: primeiramente realizamos uma simulação de Monte Carlo em um modelo GAMLSS (com uma distribuição especifica) e com oito covariáveis geradas usando a estratégia B como procedimento de seleção de variáveis explicativas para selecionar as variáveis corretas para o modelo. Com o resultado obtido com a simulação usando a estratégia B, foi utilizado somente as variáveis selecionadas e realiza-se uma nova simulação usando como procedimento de seleção a estratégia A. Para este cenário em cada simulação, adotou-se três tamanhos amostrais (150, 300 e 500) e 1000 réplicas de Monte Carlo.

4.3.1 Resultados do estudo de simulação de Monte Carlo para selecionar variáveis explicativas usando a Estratégia C em modelos com estrutura linear e dados normais.

Neste cenário, adotadou-se apenas modelos com um estrutura linear, seguindo o mesmo esquema proposto na seção 4.1.1, ou seja, consideramos uma variável aleatória $Y \sim N(\mu, \sigma)$ e foi utilizado a mesma estrutura de regressão proposta por Ramires et al. (2021).

A Tabela 9 mostra os percentuais de vezes que cada covariável foi selecionada corretamente no estudo de simulação de Monte Carlo, usando como procedimento de seleção de variáveis explicativa Estratégia C.

Para o parâmetro μ, os resultados, quando comparamos com os resultados obtidos no estudo de simulação de Monte Carlos usando o procedimento de seleção Estratégia A apresentados na Tabela 1, percebe-se que os resultados apresentados para a distribuição normal são similares aos resultados apresentados na Seção 4.1.1. Ao comparar os resultados apresentados, foi possível concluir que a Estratégia C para este cenário uma leve melhora nos percentuais se comparada com os resultados da Estratégia A. A mesma situação fica configurada para o parâmetro σ, ou seja, resultados semelhantes o que implica em uma mesma análise.

Tabela 9 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo Normal com estrutura lineare diferentes valores para κ.

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	100,00	18,70	96,70	15,40	63,10	14,90	76,00	17,70
300	μ	100,00	15,40	100,00	14,40	84,40	15,10	93,70	16,40
500	μ	100,00	15,00	100,00	14,30	96,40	15,80	99,50	15,50
150	σ	19,30	100,00	92,30	17,60	18,90	18,50	18,90	18,70
300	σ	17,10	100,00	99,30	16,20	16,30	20,00	19,20	15,50
500	σ	18,60	100,00	100,00	17,20	17,50	17,50	15,20	16,20
				Para $\kappa = [(log(n) + 2)/2]$, (Critério BIC)					
150	μ	100,00	3,80	86,30	3,10	32,10	2,20	45,70	2,60
300	μ	100,00	1,30	98,80	0,90	54,90	1,60	70,40	1,70
500	μ	100,00	1,70	100,00	1,00	76,60	1,60	91,10	1,10
150	σ	4,20	99,80	74,80	2,90	3,60	3,70	3,70	4,50
300	σ	2,50	100,00	96,10	1,70	2,10	2,30	2,20	2,60
500	σ	1,10	100,00	99,90	2,10	1,20	1,40	1,60	1,70
				Para $\kappa = [(log(n) + 2)/2]$					
150	μ	100,00	7,70	92,50	7,00	46,40	5,20	59,70	6,80
300	μ	100,00	5,00	99,70	4,50	69,10	4,40	83,90	5,10
500	μ	100,00	3,90	100,00	4,00	88,40	3,80	96,70	4,30
150	σ	8,70	99,90	83,80	7,60	9,00	9,10	7,80	8,00
300	σ	6,20	100,00	98,30	5,40	5,10	6,40	6,30	5,90
500	σ	5,80	100,00	99,90	5,80	4,40	5,20	4,80	4,40

Fonte: Autoria própria (2023)

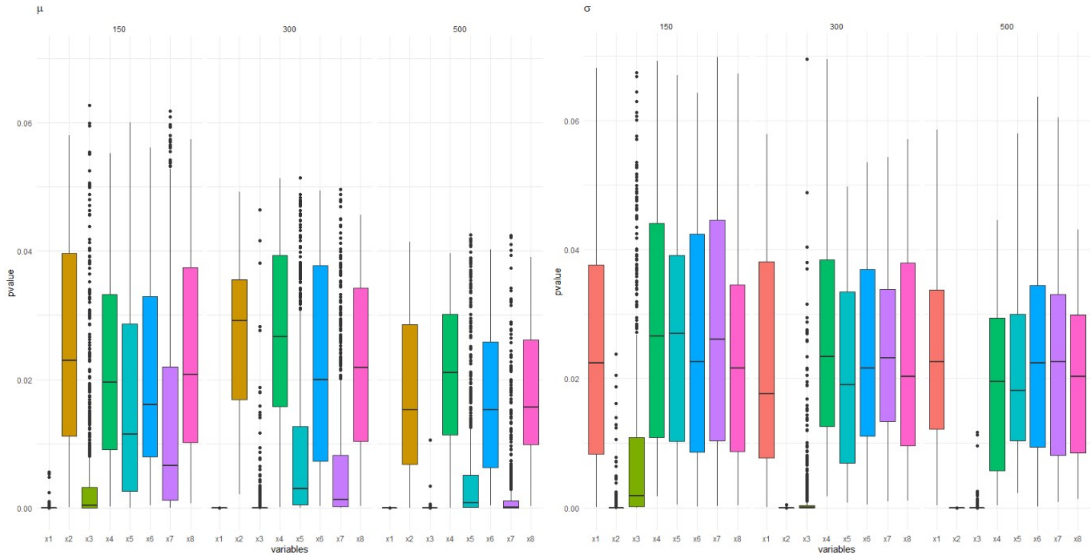
Para o parâmetro σ, os resultados, ao comparar com os resultados obtidos na Tabela 9, podemos perceber que os resultados obtidos são semelhantes aos resultados obtidos na Tabela 9, ou seja, as variáveis x_6 e x_7 foram selecionadas poucas vezes quando de fato se esperava um alto índice de seleção, dado que as mesmas estão no submodelo para σ no estudo de simulação de Monte Carlos usando o procedimento de seleção Estratégia A apresentados na Tabela 1, são semelhantes aos resultados obtidos na Tabela 9, sendo que a estratégia C apresentou pequena melhora nos seus índices.

Um outro cenário considerado no processo de seleção de variáveis explicativas, foi calcular

o p-valor de cada variável considerada no estudo, em cada réplica. Os resultados são mostrados nos Boxplot 32, 31 e 32 que apresentam os p-valores para cada variável explicativa em cada uma da três penalizações usadas. De acordo com os resultados, podemos perceber que para o submodelo para μ, as variáveis x_1 e x_3 contidas no modelo obtiveram p-valores muito baixo, muito próximo de zero, ou seja, bons valores. Já as variáveis x_5 e x_7 obtiveram p-valores bem maiores, muitos deles atípicos, mas com uma tendência de queda quando o tamanho da amostra aumenta.

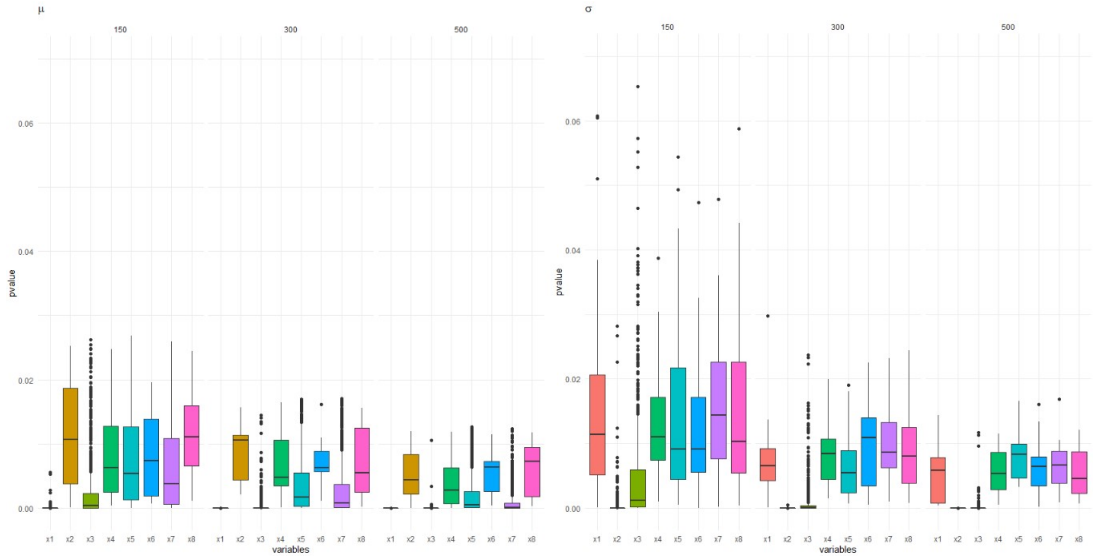
Situação semelhante acontece com as variáveis explicativas contidas no submodelo para σ, as variáveis x_2 e x_3 tem baixos p-valores, embora muitos valores atípicos e as variáveis explicativas x_6 e x_7 com os p-valores bem mais dispersos e maiores, porém sem nenhuma tendência de queda de valores quando diminuímos o tamanho amostral. Dos três critérios utilizados, o BIC apresentou os menores p-valores e a menor variabilidade.

Figura 30 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo Normal com estrutura não linear e com penalidade $\kappa = 2$ (AIC)



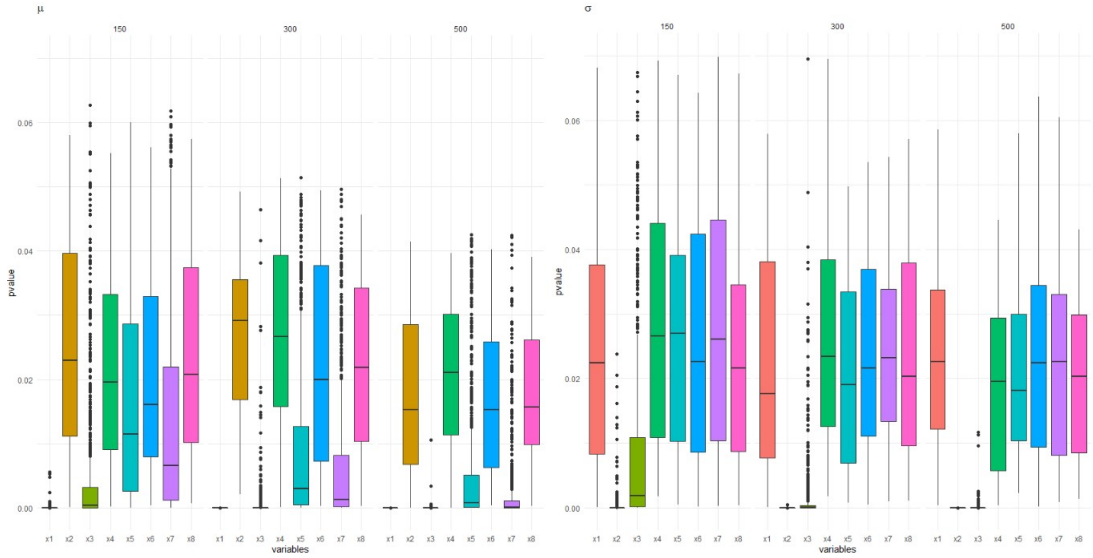
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 31 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo Normal com estrutura não linear e com penalidade $\kappa = 2$ (BIC)



Fonte: Adaptado de (RAMIRES et al., 2021)

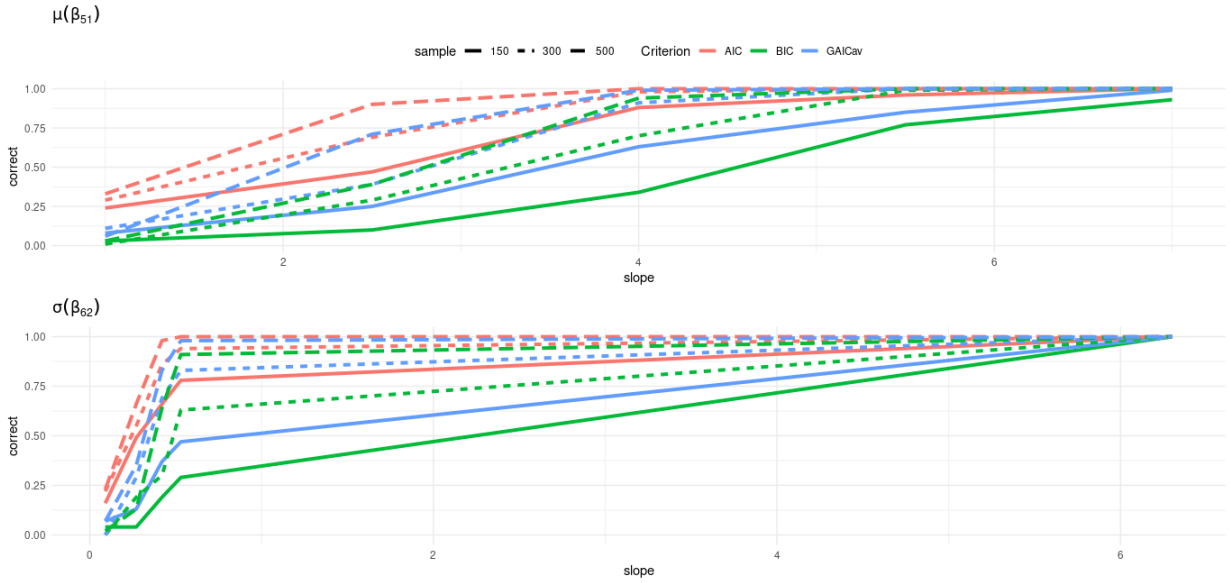
Figura 32 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo Normal com estrutura não linear e com penalidade $\kappa = [(\log(n) + 2)/2]$



Fonte: Adaptado de (RAMIRES et al., 2021)

O Gráfico 33 apresenta o desempenho dos critérios de seleção de modelos quando usados nos procedimentos de seleção de variáveis explicativas considerando a inclinação do modelo em questão. Os resultados mostram que o critério AIC apresenta maior taxa de seleção correta. Além disso foi possível concluir que o aumento na inclinação do modelo implica no crescimento da proporção de acerto do critério.

Figura 33 – Taxa de acertos da seleção de variáveis usando a Estratégia C com modelo NORMAL, diferentes inclinações e diferentes penalidades.



Fonte: Adaptado de (RAMIRES et al., 2021)

4.3.2 Estudo de simulação de Monte Carlo para selecionar variáveis explicativas usando a Estratégia C em modelos com estrutura linear e dados de contagem.

Aqui foi adotado um modelo com estrutura linear, ou seja, consideramos uma variável aleatória $Y \sim ZIP(\mu, \sigma)$, seguindo a mesma proposta descrita em Ramires et al. (2021).

Para este cenário, a Tabela 10 estão presentes os percentuais de variáveis explicativas selecionada corretamente usando a Estratégia C para o modelo ZIP. Para o submodelos do parâmetro μ, das variáveis contidas no modelo, apenas a variável x_1 foi bem selecionada, principalmente para o critério AIC, já as demais (x_3, x_5 e x_7) obtiveram os piores resultados, não passando de 21% em todos os três tipos de penalidades. Importante destacar neste cenário é o fato de que, com exceção da variável explicativa x_1 todas as demais variáveis (autenticas ou de ruído) obtiveram resultados ruins, ou seja, os percentuais de seleção correta para as variáveis autênticas.

Para o parâmetro σ, que tem como variáveis autenticas (x_2, x_3, x_6 e x_7) apenas as variáveis x_6 e x_7 tiveram baixos resultados.

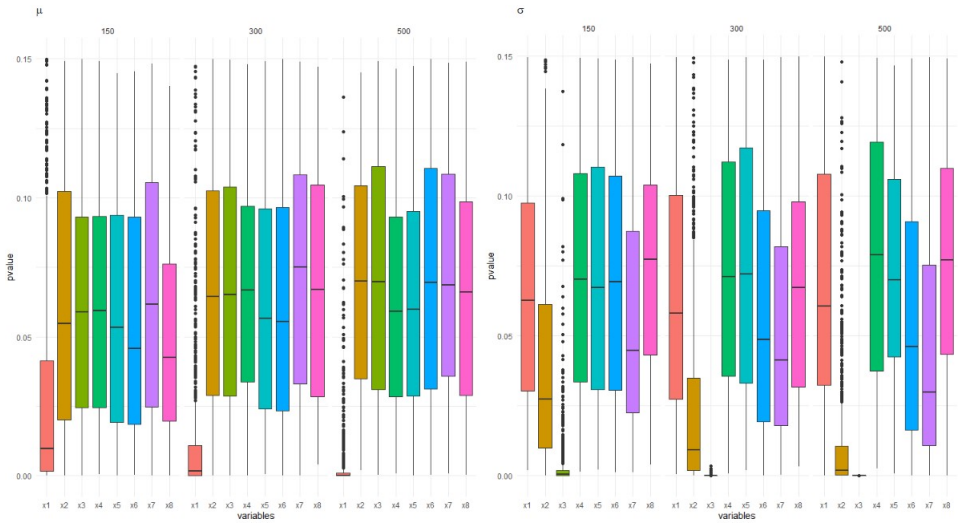
Tabela 10 – Resultados da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo ZIP com estrutura linear e diferentes valores para κ

n	Parâmetro	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
				Para $\kappa = 2$, (Critério AIC)					
150	μ	79,10	20,00	20,40	14,60	17,30	15,50	22,00	14,40
300	μ	94,90	18,00	18,30	14,90	14,00	13,00	20,70	101,70
500	μ	99,40	19,20	21,00	13,80	14,60	14,30	20,30	15,00
150	σ	20,70	58,00	99,00	16,90	16,30	22,90	29,50	17,70
300	σ	17,00	81,70	100,00	14,30	14,90	29,90	38,90	17,00
500	σ	18,50	92,10	100,00	14,90	17,00	40,20	50,40	13,80
				Para $\kappa = log(n)$, (Critério BIC)					
150	μ	52,70	5,80	4,80	1,90	2,50	1,80	3,10	1,90
300	μ	73,70	5,50	2,80	1,50	1,30	1,90	2,80	1,30
500	μ	94,50	3,10	2,10	0,70	0,70	2,00	2,30	1,20
150	σ	3,90	22,60	97,80	2,70	2,20	4,90	8,40	3,30
300	σ	2,20	43,90	100,00	1,90	1,00	6,50	10,10	1,30
500	σ	1,60	65,10	100,00	1,00	1,20	8,80	15,70	0,90
				Para $\kappa = [(log(n) + 2)/2]$					
150	μ	63,10	10,30	9,30	5,60	4,80	5,60	8,60	7,20
300	μ	84,80	9,50	7,00	4,80	3,70	5,40	6,70	5,20
500	μ	97,30	6,50	7,70	4,60	4,20	4,90	5,70	5,30
150	σ	8,50	40,30	97,80	6,10	7,00	11,00	19,40	7,00
300	σ	5,60	62,80	100,00	5,00	4,70	14,70	19,50	5,60
500	σ	4,40	80,90	100,00	3,70	4,60	19,70	28,40	4,30

Fonte: Autoria própria (2023)

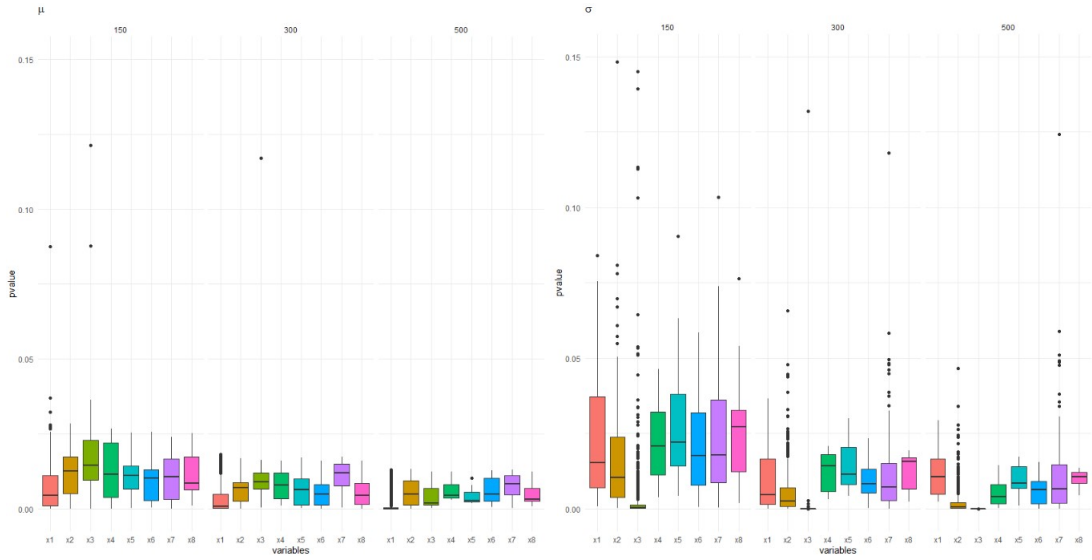
Aqui, os Gráficos Boxplot 34, 35 e 36 estão presentes os p-valores calculados no estudo de simulação para cada uma das variáveis consideradas no estudo. De acordo com os resultados mostrados, podemos perceber que o Critério AIC os mesmos apresentam uma maior variabilidade, enquanto que para os demais critérios utlizados os graficos mostram uma variação muito pequena.

Figura 34 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo ZIP com estrutura não linear e com penalidade $\kappa = 2$ (AIC)



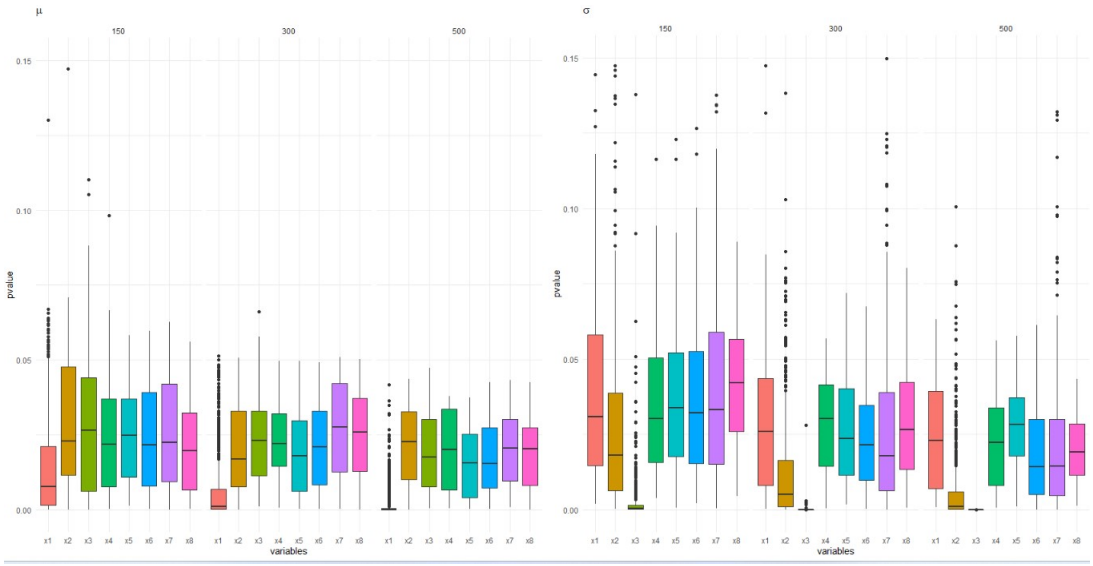
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 35 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo ZIP com estrutura não linear e com penalidade $\kappa = log(n)$ (BIC)



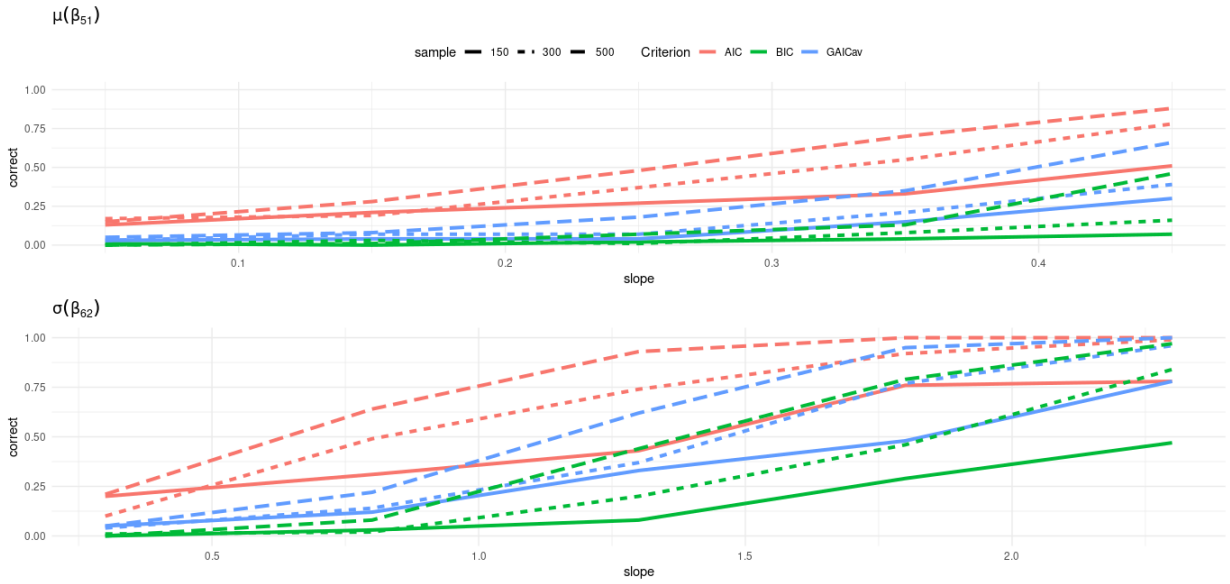
Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 36 – P-valores da seleção de variáveis explicativas usando o procedimento Estratégia C em um modelo ZIP com estrutura não linear e com penalidade $\kappa = [(log(n) + 2)/2]$



Fonte: Adaptado de (RAMIRES et al., 2021)

Figura 37 – Taxa de acertos da seleção de variáveis usando a Estratégia C com modelo ZIP, diferentes inclinações e diferentes penalidades.



Fonte: Adaptado de (RAMIRES et al., 2021)

5 APLICAÇÃO A DADOS REAIS

A aplicação terá como foco estudar e comparar técnicas de seleção de variáveis nos modelos GAMLSS. Assim, é apresentado uma base de dados com informações de uma pesquisa que foi realizada em abril de 1993 pela Infratest Sozialforschung. Este conjunto de dados estão disponíveis no pacote **gamlss.data**, chamados de `rent`, e nele encontram-se informações de uma amostra aleatória de acomodações com novos contratos de locação ou aumentos de aluguel nos últimos quatro anos em Munique - Alemanha e inclui: i) quartos individuais, ii) pequenos apartamentos, iii) apartamentos, iv) casas para duas famílias. Os alojamentos sujeitos a rendas de controle de preços, as casas unifamiliares e as casas especiais, como as coberturas, foram excluídas por serem bastante diferente das restantes e por serem consideradas um mercado distinto.

Este banco de dados é composto de 1967 observações, com 9 variáveis sendo considerada como variável resposta o valor do aluguel **R**. Para este estudo consideramos 6 variáveis explicativas apropriadas para esta abordagem.

Segue uma breve descrição das **6** variáveis que foram consideradas neste estudo:

1. **R** é a variável de resposta ao aluguel, o aluguel líquido mensal em DM, ou seja, o aluguel mensal menos o custo calculado ou estimado da utilidade:

2. **Fl** é área útil em metros quadrados;

3. **A** é o ano de construção;

4. **B** é um fator com níveis: **B1** se há banheiro, 1, (1925 obs.) ou **B2** não, 0, (44 obs.)

5. **H** é um fator com níveis: **H1** se há aquecimento central, 1, (1580 obs.) ou **H2** não, 0, (389 obs.);

6. **L**: fator com níveis: **L1** se o equipamento da cozinha está acima da média, 1, (161 obs.) ou **L2** não, 0, (1808 obs);

7. **Loc**: é um fator (composto pela combinação de **SP** uma variável que indica se o local está acima da média e **Sm** uma variável que indica se a localização está abaixo da média ou não) com três níveis que indica se a localização está: **Loc1** abaixo da média, **Loc2** na média, ou **Loc3** acima da média.

5.1 SELEÇÃO DE VARIÁVEIS EXPLICATIVAS USANDO OS PROCEDIMENTOS ESTRATÉGIA A, ESTRATÉGIA B E ESTRATÉGIA C, COM DISTRIBUIÇÃO NORMAL

Com vista em comparar os procedimentos de seleção de variáveis explicativas proposto por Stasinopoulos et al. (2017). Aqui, foi modelado o valor liquido do aluguel com base em 6 variáveis que possivelmente tenha alguma relação com o valor do aluguel, para tal, usando um modelo GAMLSS com a distribuição Normal e usando a função de ligação identidade para o parâmetro μ e a função de ligação logarítmica para o parâmetro σ. Assim, o modelo estatistico pode ser definido como:

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6$$
$$\sigma = exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6\}$$

Consequentemente o modelo em estudo é descrito pelas equações:

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1.Fl + \hat{\beta}_2.A + \hat{\beta}_3.B + \hat{\beta}_4.H + \hat{\beta}_5.L + \hat{\beta}_6.Loc$$
$$\hat{\sigma} = exp\{\hat{\beta}_0 + \hat{\beta}_1.Fl + \hat{\beta}_2.A + \hat{\beta}_3.B + \hat{\beta}_4.H + \hat{\beta}_5.L + \hat{\beta}_6.Loc\}$$

5.1.1 Cenário 1 - Seleção de variáveis explicativas sem o uso de funções de suavização, usando a distribuição normal aplicados a dados reais

A Tabela 11 estão presentes os resultados da seleção de variáveis explicativas usando as Estratégias A e Esgratégias B em modelos GAMLSS aplicados ao conjunto de dados **Rent**, sem o uso de funções de suavizações. Com os resultados obtidos, podemos perceber que para as duas estratégias (A e B) selecionaram todas as variáveis consideradas disponíveis no conjunto de dados, mas para as variáveis categóricas **B**, **H**, **L** e **Loc**, seus fatores negativos ("não há banheiros", "não há aquecimento central", "equipamento de cozinha abaixo da média"e "localização abaixo da média") não foram selecionados respectivamente em nenhuma das estratégias, o que de fato faz sentido , pois esses fatores impactariam negativamente no valor do aluguel (variável **R**).

Para modelos ajustados com as estratégias A e B no final do processo de seleção, como são as mesmas variáveis, foi possível perceber que os modelos adquiridos depois do uso da Estratégia A obteve menor AIC, portanto o melhor modelo.

Tabela 11 – Resultados da seleção de variáveis explicativas usando como métodos a Estratégia A, Estratégia B e Estratégia C, para distribuição Normal sem o uso de funções de suavização.

Parâmetro	Variáveis explicativas selecionadas											AIC
	Fl	A	B		H		L		Loc			
			B1	B2	H1	H2	L1	L2	Loc1	Loc2	Loc3	
Estratégia A												
μ	x	x	x		x		x			x	x	27858,82
σ	x	x	x		x		x			x	x	
Estratégia B												
μ e σ	x	x	x		x		x			x	x	27859,31
Estratégia C												
μ	x	x	x		x		x			x	x	27858,82
σ	x	x	x		x		x			x	x	

Fonte: Autoria própria (2023)

Quanto a Estratégia C, a análise e interpretação dos resultados deve ser a mesma observada para a Estratégia A. Esse fato é devido aos resultados apresentados na Tabela 11 que mostra que a Estratégia A e B selecionaram as mesmas variáveis, consequentemente a Estratégia C terá os mesmos resultados, pois a mesma é usa as variáveis selecionadas pela Estratégia B usa o procedimento Estratégia A. Quanto ao tempo computacional usado na Estratégia C, o mesmo é equivalente à Estratégia A.

5.1.2 Cenário 2 - Seleção de variáveis explicativas com o uso de funções de suavização, usando a distribuição normal aplicados a dados reais

Ajustou-se o modelo, usando a estratégia A, Estratégia B e Estratégia C e considerando o valor do aluguel como variável resposta e foi utilizado funções de suavização (`p-splines`) em duas variáveis explicativas (**FL** e **A**).

Os resultados apresentados na Tabela 12 mostra as variáveis que foram selecionadas para o modelo usando como método de seleção de variáveis as Estratégia A, B e C, para os submodelos μ e σ, com distribuição Normal. Para este cenário foram utilizadas suavizadores `p-splines` para as variáveis **FL** e **A** Podemos perceber que a Estratégia A, para o submodelo σ, a variável **Loc**, que indica se o local está abaixo da média, na média ou acima da média, só selecionou o fator "na média", mas para o submodelo μ, para a variável **Loc**, os fatores na média e acima

da média foi selecionado. Observando os resultados obtidos pela Estratégia B, percebemos que para a variável **Loc**, foi selecionado os fatores "na média"e "acima da média". Comparando os modelos finais após os ajustes do processo de seleção, podemos concluir que a Estratégia A obteve o melhor modelo, pois o AIC é menor quando comparamos com a Estratégia B.

Ainda de acordo com a os resultados obtidos na Tabela 12 a Estratégia C selecionou as mesmas variáveis obtidas na Estratégia B e portanto uma mesma análise e interpretação. Uma justificativa para estes resultados pode ser um número pequeno de variáveis contidas no conjunto de dados. Aqui, com o uso de `p-splines`, o custo computacional foi um pouco maior, mas semelhante á Estratégia A.

Tabela 12 – Resultados da seleção de variáveis usando métodos Estratégias A, B e C, para distribuição Normal com funções suavizadoras (`p-spline`) e parâmetros μ e σ.

Parâmetro	Variáveis explicativas selecionadas											AIC
	Fl	A	B		H		L		Loc			
			B1	B2	H1	H2	L1	L2	Loc1	Loc2	Loc3	
Estratégia A												
μ	x	x	x		x		x			x	x	27740,49
σ	x	x	x		x		x			x		
Estratégia B												
μ e σ	x	x	x		x		x			x	x	27740,83
Estratégia C												
μ	x	x	x		x		x			x	x	27740,94
σ	x	x	x		x		x			x		

Fonte: Autoria própria (2023)

Quando comparamos os resultados obtidos na seleção de variáveis quando foi usado função suavizado com os resultados obtidos quando usados funções suavizadora, podemos concluir que o modelo com funções suavizadoras seleciona menos variável e os seus respectivos modelos finais são melhores, pois os valores dos AICs são menores. O valor do AIC nas Estratégia A e C é igual, mas isso se justifica pelo uso do mesmo conjunto de variáveis explicativas.

5.2 SELEÇÃO DE VARIÁVEIS EXPLICATIVAS USANDO OS PROCEDIMENTOS ESTRATÉGIA A, ESTRATÉGIA B E ESTRATÉGIA C, COM DISTRIBUIÇÃO BCCGO

Aqui foi considerado a distribuição de probabilidade BCCGo com vista em comparar os procedimentos de seleção de variáveis explicativas proposto por Stasinopoulos et al. (2017), usando o mesmo conjunto de dados (**Rent**) que modela o valor liquido do aluguel com base em 6 variáveis que possivelmente tenha alguma relação com o valor do aluguel, para tal, usando um modelo GAMLSS com a distribuição BCCGo, uma distribuição de três parâmetros. Se $Y \sim BCCGo(\mu, \sigma, \nu)$, com $-0 < y < \infty$, $-0 < \mu < \infty$ o parâmetro de locação, $\sigma > 0$ o parâmetro de escala e $-\infty < \nu < \infty$ o parâmetro de assimetria e as devidas funções de ligação.

Assim, o modelo estatístico pode ser definido como:

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6$$

$$\sigma = exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6\} \nu = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6$$

Consequentemente o modelo em estudo é descrito pelas equações:

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1.Fl + \hat{\beta}_2.A + \hat{\beta}_3.B + \hat{\beta}_4.H + \hat{\beta}_5.L + \hat{\beta}_6.Loc$$

$$\hat{\sigma} = exp\{\hat{\beta}_0 + \hat{\beta}_1.Fl + \hat{\beta}_2.A + \hat{\beta}_3.B + \hat{\beta}_4.H + \hat{\beta}_5.L + \hat{\beta}_6.Loc\}$$

$$\hat{\nu} = \hat{\beta}_0 + \hat{\beta}_1.Fl + \hat{\beta}_2.A + \hat{\beta}_3.B + \hat{\beta}_4.H + \hat{\beta}_5.L + \hat{\beta}_6.Loc$$

5.2.1 Cenário 1 - Seleção de variáveis explicativas sem o uso de funções de suavização, usando a distribuição BCCGo aplicados a dados reais

A Tabela 5.2.1 apresenta os resultados da seleção de variáveis explicativas usando as Estratégias A, B e C em modelos GAMLSS aplicados ao conjunto de dados **Rent**. Com os resultados obtidos, percebe-se que para o parâmetro μ, todas as variáveis estratégias (A, B e C) selecionaram todas as variáveis explicativas, mas não selecionaram os fatores negativos (ou seja, aqueles fatores que não afetam positivamente a variável resposta, ou seja, não afetam positivamente no valor do aluguel. Para os parâmetros σ e ν as estratégias A e C selecionaram apenas duas variáveis explicativas. Já a Estratégia B, selecionou todas as variáveis

e consequentemente estas variáveis serão para todos os três parâmetros do modelo, o que pode ser um problema.

Os modelos ajustados no final do processo de seleção obtiveram os mesmos valores para o AIC.

Tabela 13 – Resultados da seleção de variáveis usando métodos Estratégias A, B, para distribuição BCCGo sem o uso de funções suavizadoras e com os parâmetros μ, σ e ν.

Parâmetro	**Variáveis explicativas selecionadas**											**AIC**
	Fl	**A**	**B**		**H**		**L**		**Loc**			
			B1	**B2**	**H1**	**H2**	**L1**	**L2**	**Loc1**	**Loc2**	**Loc3**	
Estratégia A												
μ	**x**	**x**	**x**		**x**		**x**			**x**	**x**	$27731,36$
σ		**x**								**x**	**x**	$27693,91$
ν		**x**										$27690,53$
Estratégia B												
μ, σ **e** ν	**x**	**x**	**x**		**x**		**x**			**x**	**x**	$27584,13$
Estratégia C												
μ	**x**	**x**	**x**		**x**		**x**			**x**	**x**	$27731,36$
σ		**x**								**x**	**x**	$27693,91$
ν		**x**										$27690,53$

Fonte: Autoria própria (2023)

5.2.2 Cenário 2 - Seleção de variáveis explicativas com o uso de funções de suavização, usando a distribuição BCCGo aplicados a dados reais

Os resultados obtidos quando usado funções de suavizações são apresentados na Tabela 5.2.2 e mostram o mesmo cenário da aplicação quando não foi usado funções de suavizações e portanto a análise e interpretação é a mesma.

Tabela 14 – Resultados da seleção de variáveis usando métodos Estratégias A, B, para distribuição BCCGo com o uso de funções suavizadoras e com os parâmetros μ, σ e ν.

Parâmetro	**Variáveis explicativas selecionadas**											**AIC**
	Fl	**A**	**B**		**H**		**L**		**Loc**			
			B1	**B2**	**H1**	**H2**	**L1**	**L2**	**Loc1**	**Loc2**	**Loc3**	
Estratégia A												
μ	x	x	x		x		x			x	x	$27571,84$
σ	x	x								x	x	$27578,54$
nu		x			x							$27690,53$
Estratégia B												
μ, σ **e** ν	x	x	x		x		x			x	x	$27584,13$
Estratégia C												
μ	x	x	x		x		x			x	x	$27571,84$
σ	x	x								x	x	$27578,54$
ν		x			x							$27690,53$

Fonte: Autoria própria (2023)

Comparando as estratégias de seleção de variáveis explicativas para a distribuição BCCGo apresentados nas Tabelas 5.2.1 e 5.2.2 pode-se concluir que a Estratégia A obteve melhores resultados.

6 CONCLUSÃO

Neste pesquisa dissertativa, foi apresentado e discutido um estudo dos métodos de seleção de variáveis explicativas em modelos GAMLSS, onde foi generalizado o método StepGAIC, especificamente as **estratégias A e B**. Foi possível concluir que no primeiro estudo de simulação de Monte Carlo realizado tem tinha o objetivo de comparar o desempenho dos dois processos de seleção de variáveis explicativas (Estratégias A e B), o procedimento Estratégia A apresentou melhores percentuais de seleção correta (para as variáveis contidas no modelo) e menores percentuais de seleção incorretas (variáveis que não estavam contidas no modelo) para aquelas variáveis que não deveriam ser selecionadas. No segundo estudo de simulação, onde consideramos os resultados obtidos da estratégia B, e usamos a estratégia A para fazer fazer uma nova seleção, os resultados foram semelhante aos resultados obtidos pela Estratégia A. Na aplicação com dados reais, foram confirmados os resultados obtidos nos estudos de simulação, ou seja, a Estratégia A e a Estratégia C tem melhor desempenho que a estratégia B, porém com esse estudo não foi possível concluir que a combinação da Estratégia B com a Estratégia A (Estratégia C) é melhor que a Estratégia A.

Para pesquisas futuros seria importante estudar o tempo computacional de cada uma das estratégias e buscar uma combinação entre qualidade de seleção de variáveis e tempo computacional. Especialmente quando se considera funções de suavização pois neste caso os procedimentos de seleção de modelos apresentados demoraram mais para mostrar os resultados.

REFERÊNCIAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: _____. *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York, 1973. p. 199–213.

ALVES, M. F.; LOTUFO, A. D. P.; LOPES, M. L. M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, v. 1, n. 1, 2013.

BAYER, F. M.; CRIBARI-NETO, F. Bootstrap-based model selection criteria for beta regressions. *Test*, Springer, v. 24, n. 4, p. 776–795, 2015.

BAYER, F. M.; CRIBARI-NETO, F. Model selection criteria in beta regression with varying dispersion. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 46, n. 1, p. 729–746, 2017.

CHARNET, R.; FREIRE, C. d. L.; CHARNET, E. M.; BONVINO, H. et al. Análise de modelos de regressão linear com aplicações. *Campinas: Unicamp*, 1999.

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

De Bastiani, F.; RIGBY, R. A.; STASINOPOULOUS, D. M.; CYSNEIROS, A. H.; URIBE-OPAZO, M. A. Gaussian markov random field spatial models in gamlss. *Journal of Applied Statistics*, Taylor & Francis, v. 45, n. 1, p. 168–186, 2018.

EFROYMSON, M. Multiple regression analysis. *Mathematical methods for digital computers*, John Wiley & Sons, p. 191–203, 1960.

EILERS, P. H.; MARX, B. D. Flexible smoothing with b-splines and penalties. *Statistical science*, Institute of Mathematical Statistics, v. 11, n. 2, p. 89–121, 1996.

EILERS, P. H.; MARX, B. D. *Practical smoothing: The joys of P-splines*. [S.l.]: Cambridge University Press, 2021.

FLORENCIO, L. de A. *Engenharia de avaliações com base em modelos GAMLSS*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2010.

HASTIE, T.; TIBSHIRANI, R. *Generalized additive models (Chapman Hall, London, UK)*. 1990.

HOCKING, R. R. The analysis and selection of variables in linear regression. *Biometrics*, JSTOR, p. 1–49, 1976.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970.

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

MILLER, A. J. *Subset selection in regression*. [S.l.]: Chapman and Hall, 1990.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. *Rev. bras. ciênc. saúde*, p. 289–310, 2008.

RAMIRES, T. G.; NAKAMURA, L. R.; RIGHETTO, A. J.; PESCIM, R. R.; MAZUCHELI, J.; RIGBY, R. A.; STASINOPOULOS, D. M. Validation of stepwise-based procedure in gamlss. *Journal of Data Science*, v. 19, n. 1, p. 96–110, 2021.

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

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

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

SCHWARZ, G. E. Estimating the dimension of a model. *Annals of Statistics*, v. 6, n. 2, p. 461–464, 1978.

SHIBATA, R. Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, Oxford University Press, v. 71, n. 1, p. 43–49, 1984.

SMITH, G. Step away from stepwise. *Journal of Big Data*, Springer, v. 5, n. 1, p. 1–12, 2018.

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

STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. D. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: CRC Press, 2017.

THOMPSON, B. *Why wont stepwise methods die?* [S.l.]: Taylor & Francis, 1989.

THOMSON, R. E.; EMERY, W. J. *Data analysis methods in physical oceanography*. [S.l.]: Newnes, 2014.

VENABLES, W.; RIPLEY, B. *Modern Applied Statistics with S Fourth edition by, World*. 2002.

VOUDOURIS, V.; GILCHRIST, R.; RIGBY, R.; SEDGWICK, J.; STASINOPOULOS, D. Modelling skewness and kurtosis with the bcpe density in gamlss. *Journal of Applied Statistics*, Taylor & Francis, v. 39, n. 6, p. 1279–1293, 2012.

WOOD, S. N. *Generalized additive models: an introduction with R*. [S.l.]: chapman and hall/CRC, 2006.

YAMASHITA, T.; YAMASHITA, K.; KAMIMURA, R. A stepwise aic method for variable selection in linear regression. *Communications in Statistics—Theory and Methods*, Taylor & Francis, v. 36, n. 13, p. 2395–2403, 2007.

ZHANG, P. Inference after variable selection in linear regression models. *Biometrika*, Oxford University Press, v. 79, n. 4, p. 741–746, 1992.