

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**Modelagem Composicional com Regressão Dirichlet: Correção
de Viés por Bootstrap e Teste para Seleção de Modelos
Não-encaixados**

Dissertação de Mestrado

Jesy Karolayne Sales dos Santos



São Cristóvão – Sergipe

2026

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Jesy Karolayne Sales dos Santos

**Modelagem Composicional com Regressão Dirichlet: Correção
de Viés por Bootstrap e Teste para Seleção de Modelos
Não-encaixados**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Kalil Araujo Bispo
Coorientador(a): Luiz Henrique Dore

São Cristóvão – Sergipe

2026

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

S237p Santos, Jesy Karolayne Sales dos
Modelagem composicional com regressão Dirichlet: correção de viés por *bootstrap* e teste para seleção de modelos não-encaixados / Jesy Karolayne Sales dos Santos ; orientador Kalil Araujo Bispo. - São Cristóvão, 2026.
106 f.; il.

Dissertação (mestrado em Ciência da Computação) –
Universidade Federal de Sergipe, 2026.

1. Computação. 2. Processo estocástico. 3. Teste de hipóteses (Estatística). 4. Modelos matemáticos. I. Bispo, Kalil Araujo orient. II. Título.

CDU 004

Agradecimentos

Ao professor Kalil pela supervisão do desenvolvimento desta dissertação e pelas orientações gerais ao longo do curso, fundamentais para a condução adequada das atividades acadêmicas. De maneira especial, gratidão ao professor Luiz Henrique, sua atuação foi central na construção deste trabalho, pelo conhecimento transmitido e acompanhamento contínuo e qualificado durante todas as etapas da pesquisa.

Ao grupo de pesquisa Reconstrução das Capturas de Pescado do Brasil pela oportunidade de integração ao projeto, pelo ambiente colaborativo e pelas discussões científicas que contribuíram para o amadurecimento das ideias e produções acadêmicas. Em especial, à professora coordenadora Kátia Meireles, pela atenciosa condução das atividades, pela confiança depositada e pela disponibilização dos dados utilizados nesta dissertação.

Ao corpo docente e equipe administrativa do Programa de Pós-Graduação, agradeço pelo conhecimento transmitido ao longo das disciplinas e orientações prestadas durante todo o percurso acadêmico.

Por fim, agradeço à FAPES e ao Ministério da Pesca e Aquicultura pelo apoio financeiro.

Resumo

Modelos de regressão são fundamentais na análise estatística, pois permitem descrever, explicar e prever relações entre variáveis em diferentes áreas do conhecimento. Em particular, o modelo de regressão Dirichlet é adequado para dados composicionais, caracterizados por representarem partes proporcionais de um todo. A estimação dos parâmetros do modelo por máxima verossimilhança pode apresentar viés significativo em amostras reduzidas, comprometendo a qualidade inferencial, ao passo que a seleção de modelos concorrentes carece de procedimentos específicos e consolidados na literatura nesse contexto de regressão. Em vista disso, esta dissertação tem como objetivo propor melhorias metodológicas para o modelo de regressão Dirichlet, com foco na correção de viés dos estimadores por meio de técnicas Bootstrap, nas versões paramétrica e não paramétrica, e a adaptação dos testes de hipóteses J e MJ pelo princípio da interseção-união para auxiliar a seleção entre modelos não encaixados. Para avaliar o desempenho das metodologias propostas, foram realizados estudos de simulação de Monte Carlo, comparando os estimadores originais e corrigidos em termos de viés e erro quadrático médio, bem como examinando o comportamento dos testes em diferentes tamanhos amostrais. Os resultados para o estudo da correção de viés demonstram que a técnica de reamostragem bootstrap é eficaz na redução do erro de estimação, com baixo impacto na medida do erro quadrático médio, principalmente na versão paramétrica. Quanto ao teste de seleção, a adaptação do teste MJ para o contexto da regressão Dirichlet apresentou taxas de rejeição nula próximas aos níveis nominais e elevada frequência de seleção do modelo corretamente especificado.

Palavras-chave: Regressão Dirichlet. Dados composicionais. correção de viés. Seleção de Modelo. Teste de Hipótese. Modelos não-aninhados.

Abstract

This study aims to investigate the effectiveness of bias correction via the Bootstrap technique in Dirichlet regression models, and to adapt hypothesis tests for non-nested models, such as the J and MJ tests, to the Dirichlet regression context, in order to support the selection of competing models. Regression models are fundamental in statistical analysis, enabling the description and prediction of relationships between variables. In particular, the Dirichlet regression model is suitable for compositional data, which are characterised by components that represent proportional parts of a whole. Parameter estimation in this model using maximum likelihood may exhibit significant bias in small samples, compromising inferential quality. Thus, this work proposes to evaluate the performance of bias correction through the Bootstrap method, in both its parametric and non-parametric versions, by comparing it with the original estimators through Monte Carlo simulations. The results demonstrate that the resampling technique is effective in reducing bias, with a variable impact on the mean squared error. Regarding the model selection procedure, the Monte Carlo experiments demonstrated that the proposed adaptation of the MJ test, based on the choice of the maximum p-value among competing models, yielded null rejection rates close to the nominal significance levels and a high frequency of correctly selecting the true model.

Keywords: Dirichlet regression. Compositional data. Bias correction. Model selection. Hypothesis testing. Non-nested models.

Lista de ilustrações

Figura 1 – Exemplo de dados composicionais cujos componentes (gordura, proteínas e carboidratos) representam nutrientes que compõe os alimentos, tais como: soja, ervilhas, milho e trigo. Em cada alimento, os componentes (Y_1, Y_2, Y_3) somam 1 (ou 100%)	14
Figura 2 – Formas da Densidade da distribuição Dirichlet com a modificação do parâmetro de forma λ . Variáveis $X = (X_1, X_2, X_3)$ simuladas com $N = 1000$. . .	20

Lista de tabelas

Tabela 1 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Cox-Snell para o modelo linearmente especificado e sem intercepto, $p = q = 3$	36
Tabela 2 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Bootstrap Paramétrico (modelo linearmente especificado).	43
Tabela 3 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Bootstrap Não Paramétrico (modelo linearmente especificado).	45
Tabela 4 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de CoxSnell.	46
Tabela 5 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Bootstrap Paramétrico (especificação polinomial).	49
Tabela 6 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Bootstrap Não Paramétrico (especificação polinomial).	50
Tabela 7 – Proporções anuais de captura artesanal das espécies de camarão em toneladas para imputação.	51
Tabela 8 – Estimativas dos parâmetros pelos métodos EMV, e versões corrigidas por Bootstrap paramétrico e não paramétrico para o modelo ajustado às capturas de camarão sete-barbas, branco, rosa e santana da pesca artesanal (N = 39).	53
Tabela 9 – Viés e erro quadrático médio (EQM) marginais para as componentes do modelo ajustado aos dados de capturas de camarão da pesca artesanal (N = 39).	53
Tabela 10 – Estimativas dos parâmetros pelos métodos EMV, e versões corrigidas por Bootstrap paramétrico e não paramétrico para o modelo ajustado as capturas de camarão sete-barbas, branco e rosa da pesca artesanal e industrial (N = 46).	55
Tabela 11 – Viés e erro quadrático médio (EQM) marginais para as componentes do modelo ajustado aos dados de capturas de camarão da pesca artesanal e industrial (N = 46).	55
Tabela 12 – Taxas de rejeição empírica da hipótese nula (%) do procedimento bootstrap baseado no princípio da Interseção-União	68
Tabela 13 – Frequências (%) da seleção do modelo verdadeiro utilizando o procedimento bootstrap baseado no princípio da Interseção-União como critério de seleção, quando a hipótese nula não foi rejeitada.	69

Lista de abreviaturas e siglas

BC	Estimador Corrigido por Viés (Bias Corrected)
BNP	Bootstrap Não Paramétrico
BP	Bootstrap Paramétrico
EMV	Estimador de Máxima Verossimilhança
EQM	Erro Quadrático Médio
FAPES	Fundação de Apoio à Pesquisa e Extensão de Sergipe
GAMLSS	Modelos Aditivos Generalizados para Localização, Escala e Forma (Generalized Additive Models for Location, Scale and Shape)
IU	Interseção-União
J	Estatística do Teste J
LRT	Teste da Razão de Verossimilhanças (Likelihood Ratio Test)
MJ	Teste MJ (Minimum J Test)
MPA	Ministério da Pesca e Aquicultura
TED	Termo de Execução Descentralizada
UFS	Universidade Federal de Sergipe

Lista de símbolos

β	Vetor de coeficientes de regressão do modelo Dirichlet.
λ_i	Parâmetro de forma da distribuição Dirichlet.
ϕ	Parâmetro de precisão (ou concentração total) da distribuição Dirichlet.
γ	Vetor de coeficientes associado ao submodelo de precisão.
η_{ir}	Preditor linear do modelo de regressão.
ϵ	Termo de erro aleatório.
θ	Vetor de parâmetros desconhecidos do modelo.
$\hat{\theta}$	Estimador de máxima verossimilhança de θ .
$\tilde{\theta}$	Estimador corrigido de viés.
n	Tamanho da amostra.
N	Número total de observações.
$f(\cdot)$	Função geral de regressão.
$f(y \lambda)$	Função densidade da distribuição Dirichlet.
$g(\cdot)$	Função de ligação do modelo de regressão.
$\Gamma(\cdot)$	Função gama.
$L(\beta)$	Função de verossimilhança.
$\ln L(\beta)$	Log-verossimilhança.
$U(\beta)$	Função score.
$E(\cdot)$	Esperança matemática.
$Var(\cdot)$	Variância.
$Cov(\cdot)$	Covariância.
$B(\theta)$	Viés do estimador.
$B_1(\theta)$	Termo de viés de segunda ordem.
I_q	Matriz identidade de dimensão $q \times q$.

K	Matriz de informação de Fisher.
K^{-1}	Inversa da matriz de informação de Fisher.
$vec(\cdot)$	Operador que empilha colunas de uma matriz formando um vetor.
$\psi(\cdot)$	Função digama, definida como a derivada logarítmica da função gama.
$\psi'(\cdot)$	Função trigama, definida como a derivada da função digama.



**UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

**Ata da Sessão Solene de Defesa da Dissertação do Curso
de Mestrado em Ciência da Computação-UFS.**

Candidata: Jesy Karolayne Sales dos Santos

Aos 26 dias do mês de junho do ano de dois mil e vinte seis, com início às 9hs, realizou-se na Sala de Seminários do PROCC da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloísio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **Jesy Karolayne Sales dos Santos** que desenvolveu o trabalho intitulado: *“Modelagem Composicional com Regressão Dirichlet: Correção de Viés por Bootstrap e Teste para Seleção de Modelos Não-encaixados”*, sob a orientação do Prof. Dr. **Kalil Araujo Bispo**. A Sessão foi presidida pelo Prof. Dr. **Kalil Araujo Bispo** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, o Dr. **Leonardo Nogueira Matos** (PROCC/UFS) e, em seguida, Dr. **Luiz Henrique Gama Dore de Araujo** (UFS) e Dr. **Getúlio José Amorim do Amaral** (UFPE). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) Aprovada “(aprovado/reprovado)”. Atendidas as exigências da Instrução Normativa 05/2019/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 04/2021/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária “Prof. José Aloísio de Campos”, 26 de junho de 2026.

**Prof. Dr. Kalil Araujo Bispo
(PROCC/UFS)
Presidente**

**Prof. Dr. Leonardo Nogueira Matos
(PROCC/UFS)
Examinador Interno**

**Prof. Dr. Luiz Henrique Gama Dore de Araujo
(UFS)
Examinador Externo ao programa**

**Prof. Dr. Getúlio José Amorim do Amaral
(UFPE)
Examinador Externo à Instituição**

**Jesy Karolayne Sales dos Santos
Candidata**

Sumário

1	Introdução	12
1.1	Objetivos	16
1.2	Organização da Dissertação	16
2	Modelo de Regressão Dirichlet	18
2.1	Distribuição Dirichlet	19
2.2	Regressão Dirichlet	20
2.2.1	Estimador de Máxima Verossimilhança	22
3	Correção de vies em modelos de Regressão	25
3.1	Revisão	25
3.2	Correção do vies do Modelo de Regressão de Dirichlet	29
3.2.1	Vies do Estimador de Máxima Verossimilhança	29
3.2.2	Correção Analítica	31
3.2.2.1	Simulação Monte Carlo	35
3.2.3	Correção por Bootstrap	37
3.2.3.1	Estimadores Corrigidos	39
3.3	Avaliações Numéricas	40
3.3.1	Resultados do Experimento	42
3.4	Aplicação	51
3.4.1	Imputação de dados faltantes na série histórica de capturas de camarão	51
3.5	Conclusão	54
4	Seleção de Modelos Dirichlet Não-Aninhados	57
4.1	Teste J e MJ para seleção de modelos Dirichlet	61
4.1.1	Testes de seleção com Bootstrap	65
4.2	Avaliação Numérica	67
4.2.1	Resultados do Experimento	68
4.3	Conclusão	69
5	Considerações Finais	71
	Referências	74

Apêndices

79

APÊNDICE A	Implementação computacional em Linguagem R para a Correção de Viés do Modelo de Regressão Dirichlet	80
A.1	Avaliação Numérica para a Correção de Viés Analítica de CoxSnell - Modelos Lineares	80
A.1.1	Modelos sem intercepto	80
A.1.2	Modelos com intercepto	87
A.2	Avaliação Numérica por Bootstrap para a Correção de Viés - Modelos Lineares	88
A.3	Avaliação Numérica por Bootstrap para a Correção de Viés - Modelos Polinomiais	90
A.4	Correção de viés em Modelos Dirichlet para a Captura de Pesca de Camarão . .	93
APÊNDICE B	Implementação computacional em Linguagem R para a Seleção de Modelos Não-Aninhados de Regressão Dirichlet	98

1

Introdução

A generalização de modelos matemáticos e algoritmos em análises estatísticas, seja na estatística tradicional ou no aprendizado de máquina, busca encontrar expressões matemáticas ou funções que ajudem a descrever padrões nos dados coletados. Esses modelos são criados a partir de uma amostra, e tem o desafio de serem úteis também para descrever o comportamento de toda população, ou seja, que tenham a capacidade de representar adequadamente a população da qual os dados se originam. Com isso, é possível usar o modelo para realizar previsões ou obter conclusões sobre o fenômeno associado aos dados, antes mesmo de observá-lo.

Essa capacidade, no entanto, é frequentemente limitada devido às incertezas inerentes aos dados, à estrutura do modelo e aos métodos de estimação. Essas imprecisões podem ocorrer, por exemplo, devido a variações amostrais, especificações incorretas do modelo, presença de ruído nos dados e viés sistemático dos estimadores ([MAVROGIORGOS et al., 2024](#)). O desafio de quantificar e mitigar essas incertezas é um tema central na ciência estatística ([DAVISON; HINKLEY, 1997](#); [LEE; CHEN, 2025](#)).

Nesse contexto, os modelos de regressão são amplamente utilizados quando o objetivo é avaliar a relação entre as variáveis envolvidas e quantificar o grau de dependência entre elas ou o efeito que uma exerce sobre a outra. Nessa relação, as variáveis ditas independentes (ou explicativas, covariáveis ou regressores) são usadas para prever ou explicar o comportamento de uma ou mais variáveis de interesse, chamadas de dependentes ou respostas, permitindo identificar padrões, tendências e até realizar projeções ([FAHRMEIR et al., 2013](#)). Em termos simples, um modelo de regressão busca encontrar uma função matemática que melhor descreva como as mudanças em uma variável resposta estão associadas às mudanças em uma variável explicativa.

As variáveis em um estudo estatístico representam características ou quantidades observáveis no mundo real e, por isso, podem assumir diferentes tipos, como contínuas, binárias, categóricas, de contagem, temporais ou espaciais, dependendo do fenômeno que se deseja mensurar. Essa diversidade exige a escolha de modelos estatísticos que sejam apropriados tanto

para o tipo das variáveis quanto para a forma como elas se distribuem. Por exemplo, o modelo clássico de regressão linear é mais adequado quando a variável resposta y é contínua e apresenta uma distribuição aproximadamente normal, condicionada às variáveis explicativas. No entanto, modelos mais gerais de regressão são necessários, por exemplo, quando a variável resposta é binária (como presença ou ausência de uma doença), quando os efeitos das variáveis explicativas não seguem uma relação linear ou quando há fatores que introduzem variações específicas por grupo ou região que precisam ser levadas em consideração no modelo (FAHRMEIR et al., 2013).

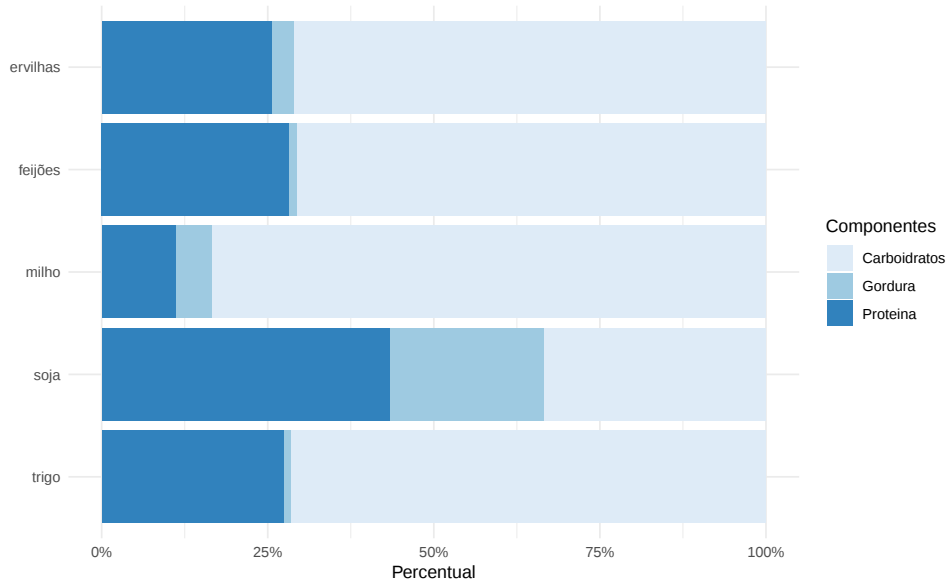
Além de considerar o tipo de variável, é importante também observar a quantidade de variáveis dependentes que estão sendo modeladas. Quando há apenas uma variável resposta, trata-se de um modelo *univariado*. Em muitas situações reais, é necessário analisar simultaneamente um conjunto de variáveis respostas relacionadas, o que caracteriza um modelo *multivariado*. Nesses casos, a proposta é modelar de forma conjunta essas variáveis, levando em conta possíveis correlações entre elas. Um exemplo de aplicação, na qual um modelo de regressão multivariado se faz necessário, é a modelagem de variáveis que constituem uma composição, como por exemplo diferentes sedimentos do solo (areia, silte, argila) ou os gastos percentuais de um orçamento, como exemplificado por Aitchison (1982).

Os dados composicionais (ou composições), são um caso específico de dados multivariados, amplamente encontrado em disciplinas como geoquímica, ciências ambientais ou biologia, sendo que cada elemento observado, dito componente, representa partes proporcionais de um todo. Estes dados, portanto, são vetores multivariados cujos componentes representam frações de um total, geralmente expressos como proporções ou porcentagens (HIJAZI; JERNIGAN, 2009). Devido a essa característica, as composições apresentam propriedades peculiares: são registradas em forma fechada, ou seja, a soma de seus componentes é constante (normalmente igual a 1 ou 100), as variáveis não assumem valores negativos e estão condicionadas a uma soma total fixa (PAWLOWSKY-GLAHN; EGOZCUE, 2006). Para ilustrar o conceito de dados composicionais, considere $Y = (Y_1, Y_2, Y_3)$ uma composição de nutrientes alimentícios, comum em tabelas nutricionais, tal que cada Y_i é um componente presente no alimento, como Y_1 = gordura, Y_2 = proteína e Y_3 = carboidratos. Este exemplo, adaptado de Boogaart e Tolosana-Delgado (2013) é apresentado na Figura 1.

A propriedade de soma constante entre os componentes condiciona as relações estatísticas entre eles, o que pode induzir a correlações espúrias (AITCHISON, 1982). Em outras palavras, mesmo que dois componentes não estejam logicamente associados, o simples fato de que o aumento em um componente exija a redução em pelo menos um outro (para manter a soma constante) pode induzir uma correlação negativa artificial (PAWLOWSKY-GLAHN; EGOZCUE, 2006). Essa característica é uma das principais dificuldades analíticas no tratamento desse tipo de dado.

Os dados composicionais ocupam um espaço amostral restrito, uma vez que as variáveis podem variar apenas de 0 a uma constante fixada. Esse espaço restrito, explorado nos trabalhos

Figura 1 – Exemplo de dados composicionais cujos componentes (gordura, proteínas e carboidratos) representam nutrientes que compõe os alimentos, tais como: soja, ervilhas, milho e trigo. Em cada alimento, os componentes (Y_1, Y_2, Y_3) somam 1 (ou 100%)



Fonte: Gráfico elaborado pelo autor com dados de composições alimentícias dada [Boogaart e Tolosana-Delgado \(2013\)](#)

de Aitchison (1982,1994), é formalmente definido como *simplex*. Para sua definição matemática, considere $x = (x_1, \dots, x_D)$ um vetor composicional com D partes, o espaço de x é dado por:

$$\mathbb{S}^d = \{(x_1, \dots, x_D) : x_i \geq 0 \ (i = 1, \dots, D), \ x_1 + \dots + x_D = 1\} \quad (1.1)$$

em que D representa o número de componentes estritamente positivas. A restrição de soma constante reduz a efetiva dimensão D das composições para $d = D - 1$, e $\mathbb{S}^d$ é dito ser o espaço amostral simplex unitário, apropriado para o estudo das composições. Dados composicionais são comumente analisados aplicando-se uma transformação que os leva do espaço $\mathbb{S}^{(D-1)}$ para o $\mathbb{R}^{D-1}$, onde métodos convencionais de análise podem ser aplicados. Certas transformações, como a transformação logística aditiva (ver [Aitchison \(1982, p. 143-144\)](#)) só podem ser utilizadas quando os dados são restritos ao interior do simplex, isto é, não possuem zeros.

Em razão das restrições e propriedades inerentes aos dados composicionais é fundamental utilizar abordagens que respeitem sua estrutura. Caso contrário, o uso de modelos inadequados podem levar a interpretações incorretas. Uma das distribuições de probabilidade, ou modelo matemático que serve para descrever o comportando desse tipo de dados multivariado é a distribuição Dirichlet, visto que ela pertence à família de distribuições definidas no espaço simplex $\mathbb{S}^d$ ([AITCHISON, 1982](#)).

Estimar os parâmetros de regressão consiste em encontrar os valores mais prováveis

desses parâmetros, de modo que o modelo seja fidedigno a realidade observada. Na estimação dos parâmetros da regressão de Dirichlet o método comumente usado é máxima verossimilhança (WICKER et al., 2008).

Entretanto, assim como em outros modelos de regressão, a regressão Dirichlet está sujeita a incertezas associadas ao processo de estimação. Em contextos de amostras pequenas ou dados assimétricos, as estimativas dos parâmetros podem ser enviesadas comprometendo a interpretação dos efeitos. Para mitigar esses erros, a literatura tem proposto diversas medidas de correção de viés em modelos de regressão, seja baseado em expressões analíticas (como as de Cox e Snell, ou Firth), ou métodos baseados em reamostragem (como jackknife e bootstrap) (CORDEIRO; CRIBARI-NETO et al., 2014). A correção de viés é, portanto, uma importante área de pesquisa, pois visa melhorar a precisão inferencial e, consequentemente, os resultados obtidos em aplicações práticas.

Outra etapa fundamental no processo de modelagem estatística refere-se à seleção de modelos. Após o ajuste de diferentes especificações para um mesmo conjunto de dados, torna-se necessário definir critérios que permitam identificar qual modelo parece ser mais adequado ao fenômeno investigado. Em aplicações reais, é comum que modelos concorrentes se diferenciem quanto aos regressores utilizados, funções de ligação adotadas ou mesmo quanto à estrutura dos submodelos associados aos parâmetros da distribuição (DAVIDSON; MACKINNON, 1981; SAPRA, 2008; CRIBARI-NETO; LUCENA, 2015).

Na literatura, a comparação entre modelos estatísticos é realizada por meio de testes de hipótese (LORESTANI; SAYYAREH, 2017) ou critérios de informação, como o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC) (SAYYAREH, 2012). Tais procedimentos são adequados apenas em situações nas quais os modelos comparados sejam aninhados (encaixados), isto é, quando um deles pode ser obtido como caso particular do outro mediante imposição de restrições sobre os parâmetros (CRIBARI-NETO; LUCENA, 2015). Nesses casos, existe uma relação hierárquica entre os modelos.

Por outro lado, modelos não encaixados são aqueles em que nenhum dos modelos pode ser reduzido ao outro. Essa situação ocorre, por exemplo, quando diferentes modelos utilizam conjuntos distintos de covariáveis, funções de ligação ou diferentes especificações para os submodelos de regressão. Conforme discutido por Cribari-Neto e Lucena (2017), quando a condição de encaixe não é satisfeita, os procedimentos usuais deixam de ser apropriados para seleção entre modelos concorrentes.

A regressão Dirichlet possui estrutura multiparamétrica, permitindo que diferentes componentes da distribuição sejam modelados simultaneamente por meio de submodelos específicos. Consequentemente, modelos concorrentes podem divergir quanto aos regressores utilizados em cada componente da composição, configurando naturalmente situações de não encaixe. Como consequência, procedimentos usuais de comparação deixam de ser válidos, tornando necessária a utilização de testes desenvolvidos para comparação entre modelos não

aninhados nesse contexto de regressão.

1.1 Objetivos

Esta dissertação tem como objetivos principais verificar a efetividade da técnica de bootstrap para correção do viés dos estimadores de máxima verossimilhança da regressão Dirichlet, e adaptar procedimentos de seleção de modelos não aninhados ao contexto dessa classe de modelos composicionais. Em particular, busca-se avaliar o desempenho do método de correção de viés baseados em reamostragem bootstrap, bem como propor uma adaptação do teste MJ para seleção de modelos concorrentes na regressão Dirichlet, considerando sua estrutura multivariada e multiparamétrica.

Para tanto, propõe-se inicialmente revisar a literatura relacionada aos principais métodos de correção de viés, tanto analíticos quanto baseados em reamostragem, aplicados aos modelos de regressão em geral, e especificamente aplicado à regressão Dirichlet, com o intuito de compreender suas fundamentações teóricas e aplicações práticas. Com o mesmo propósito, o estudo também visa realizar uma revisão sobre o método de reamostragem Bootstrap e sua utilidade para correção de viés em estimadores obtidos por máxima verossimilhança.

Por meio de simulações numéricas, o estudo propõe avaliar o comportamento dos estimadores corrigidos, considerando medidas como viés e erro quadrático médio em diferentes tamanhos amostrais. De forma complementar, pretende-se apresentar aplicações práticas dessa abordagem, usando o modelo de regressão para solução do problema de imputação de dados da reconstrução da estatística pesqueira.

Por fim, busca-se também analisar a adequabilidade do teste de seleção e sua versão bootstrap ao contexto da regressão Dirichlet por meio de simulação Monte Carlo.

1.2 Organização da Dissertação

A presente dissertação compreende cinco capítulos, sendo este primeiro dedicado a introdução do trabalho. No capítulo 2 são apresentadas as principais características e propriedades da distribuição e modelo de regressão Dirichlet, incluindo sua formulação probabilística, parametrização e o procedimento de estimação pelo método da máxima verossimilhança.

No Capítulo 3 é apresentada uma breve revisão sobre a correção de viés dos estimadores de máxima verossimilhança em modelos de regressão, abrangendo uma descrição de alguns métodos de correção usualmente encontrados na literatura. Discute-se a literatura a respeito da correção do viés no modelo de regressão Dirichlet pelo método analítico e é apresentada a proposta de correção pela técnica Bootstrap. Também são avaliadas as simulações de Monte Carlo e discutido o desempenho dos estimadores corrigidos, comparando as versões paramétricas

e não paramétrica do Bootstrap, finalizando com as conclusões específicas relacionadas ao procedimento de correção de viés.

O Capítulo 4 discute a fundamentação teórica sobre os testes de hipóteses utilizados para seleção de modelos não encaixados, incluindo o princípio da Interseção-União, com ênfase nos testes J e MJ . Em seguida, é proposta a metodologia para a adaptação dos testes ao contexto da regressão Dirichlet, considerando sua estrutura multiparamétrica e a versão bootstrap do teste. Também são apresentados estudos de simulação de Monte Carlo para avaliação do comportamento dos testes, encerrando o capítulo com conclusões específicas acerca dos resultados obtidos. Por fim, o Capítulo 5 apresenta as principais conclusões do estudo.

2

Modelo de Regressão Dirichlet

Na análise de regressão com variáveis composicionais, é comum que os dados apresentem características como assimetria e heterocedasticidade, isto é, variabilidade não constante dos dados ao longo do intervalo de observação, especialmente porque as variáveis respostas estão restritas a um intervalo limitado. Essas características contrariam suposições básicas dos modelos de regressão lineares, que assumem que os dados são bem distribuídos (de forma simétrica, como uma curva normal) e têm variabilidade constante.

Uma estratégia comum para lidar com essas características é transformar os dados usando, por exemplo, funções logarítmicas ou razão-log, para que se aproximem das condições adequadas às suposições de modelos estatísticos tradicionais ([AITCHISON, 1982](#); [HRON; TEMPL; FILZMOSER, 2010](#)). No entanto, essas transformações fazem com que os resultados do modelo não sejam interpretados diretamente nos termos originais dos dados, e ainda podem persistir com a heterocedasticidade ([MAIER, 2014](#); [HRON; TEMPL; FILZMOSER, 2010](#)).

Nesse cenário, os modelos baseados na distribuição de probabilidade multivariada Dirichlet mostram-se como uma alternativa apropriada para lidar com dados composicionais. Essa distribuição permite modelar diretamente vetores de proporções condicionados às variáveis explicativas, respeitando as restrições do espaço onde tais dados estão definidos - o simplex, espaço anteriormente definido na equação (1.1). Apesar de a distribuição Dirichlet impor restrições específicas, como a dependência estrutural entre componentes e a presença de correlações negativas naturais ([ONGARO; MIGLIORATI, 2013](#)), o modelo de regressão baseado nessa distribuição é considerado flexível e adequado para uma ampla gama de padrões de comportamento dos dados composicionais ([HIJAZI; JERNIGAN, 2009](#)), sendo inclusive a regressão Beta, que é usado para proporções em uma única variável, um caso particular desse modelo ([MELO; VASCONCELLOS; LEMONTE, 2009](#)).

As próximas seções deste capítulo concentram-se em descrever matematicamente o modelo de regressão Dirichlet, suas propriedades e o método de estimação dos parâmetros de

regressão.

2.1 Distribuição Dirichlet

Seja um vetor aleatório $Y = (Y_1, Y_2, \dots, Y_q)$ constituído por q componentes, tal que Y pertence ao simplex positivo $\mathbb{S}^q$, onde:

- $Y_i > 0$ para todo i , portanto são valores estritamente positivos;
- $\sum_{i=1}^q Y_i = 1$, logo possui soma constante.

então Y segue uma Distribuição Dirichlet com parâmetro de forma $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_q)$, escrita como $Y \sim \mathcal{D}(\lambda_1, \lambda_2, \dots, \lambda_q)$.

A função de densidade de probabilidade de Y é definida por

$$f(y_1, \dots, y_q | \lambda_1, \dots, \lambda_q) = \prod_{i=1}^q y_i^{\lambda_i-1} \frac{\Gamma(\sum_{i=1}^q \lambda_i)}{\prod_{i=1}^q \Gamma(\lambda_i)} \quad (2.1)$$

onde $\Gamma(\cdot)$ é a função gama definida como

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

O parâmetro de concentração total de $Y \sim \mathcal{D}(\lambda)$ é dado por

$$\phi = \sum_{i=1}^q \lambda_i$$

A média e variância são definidas como

$$E[Y_i] = \frac{\lambda_i}{\phi}$$

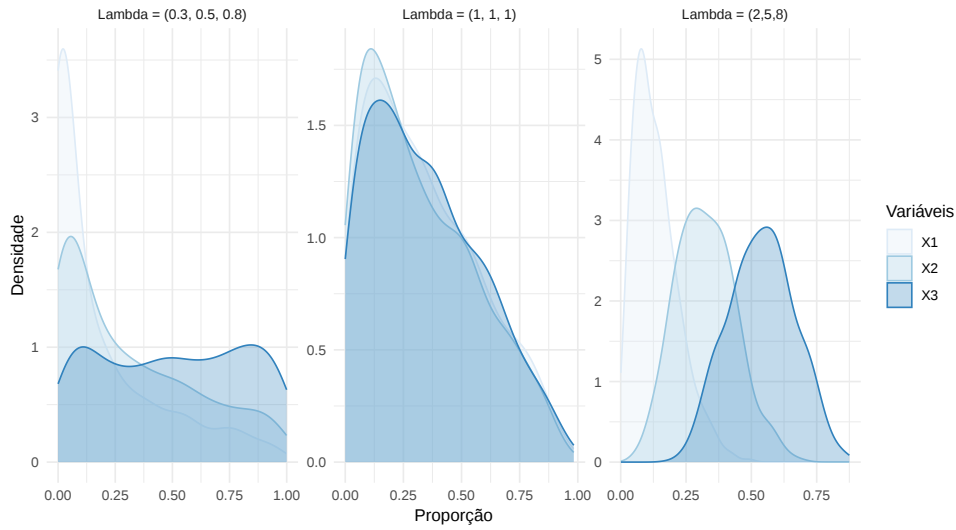
$$Var(Y_i) = \frac{\lambda_i(\phi - \lambda_i)}{\phi^2(\phi + 1)}$$

Ou seja, a média de cada componente i é a proporção de seu parâmetro λ_i em relação a soma total dos parâmetros.

A covariância entre Y_i e Y_j , para $(i, j) = 1, \dots, q$ e $i \neq j$ é calculada por

$$COV(Y_i, Y_j) = \frac{-\lambda_i \lambda_j}{\phi^2(\phi + 1)}$$

Figura 2 – Formas da Densidade da distribuição Dirichlet com a modificação do parâmetro de forma λ . Variáveis $X = (X_1, X_2, X_3)$ simuladas com $N = 1000$



A covariância da distribuição é necessariamente negativa, o que reflete a dependência entre os componentes. Assim, o aumento em uma parte resulta na redução da outra, uma vez que a soma é constante (AITCHISON, 1982).

A Figura 2 ilustra como a forma da função densidade de probabilidade da distribuição Dirichlet muda à medida que variam os valores de seu parâmetro de forma, representados por λ . Esse parâmetro controla o grau de concentração (ou dispersão) dos dados em torno da média. Valores maiores de λ resultam em distribuições mais concentradas, enquanto valores menores permitem maior dispersão e variabilidade em cada componente.

A restrição dos dados composicionais de soma constante igual a um, implica em algumas propriedades que permitem maior flexibilidade da modelagem usando a classe de distribuição Dirichlet, por exemplo, qualquer subvetor de dimensão $(q - 1)$ de Y é contínuo com a mesma função densidade dada na Equação 2.1, a permutação e a junção de componentes em um vetor menor também conduz para a mesma classe de distribuição, desde que executado a mesma permutação e agregação dos parâmetros correspondentes (MELO; VASCONCELLOS; LEMONTE, 2009).

2.2 Regressão Dirichlet

O modelo de regressão Dirichlet é uma extensão da distribuição multivariada Dirichlet que incorpora o efeito de covariáveis sobre um vetor de proporções ou composições.

Seja $Y_i^T = (Y_{i1}, \dots, Y_{iq})$ o vetor de variáveis resposta com q componentes para a i -ésima unidade amostral, para cada $i \in \{1, \dots, n\}$, e $x_i^T = (x_{i1}, \dots, x_{ip})$ o vetor de p variáveis preditoras. Um modelo de regressão Dirichlet, associando Y a x , pode ser formulado como:

- $Y_i \sim \text{Dirichlet}(\lambda_{i1}, \dots, \lambda_{iq});$
- $\eta_{ir}(x; \beta) = \beta_0 + \beta_{1r}x_{i1} + \dots + \beta_{pr}x_{ip} = x_i^T \beta_r, \forall r \in \{1, \dots, q\};$
- $g(\lambda_{ir}) = \eta_{ir}(x; \beta).$

onde $\eta_{ir}(x; \beta)$ é o preditor linear e $\beta_r^T = (\beta_{1r}, \dots, \beta_{pr})$ é o vetor de coeficientes desconhecidos da componente r . A função de ligação $g(\cdot)$ tem o objetivo de transformar a escala do parâmetros da distribuição para possibilitar que eles sejam modelados por uma função linear, e apresenta a característica de ser estritamente monótona e duas vezes diferenciável. Tal qual a função de ligação logarítmica (MORAIS; THOMAS-AGNAN; SIMIONI, 2018) ou polinomial (HIJAZI; JERNIGAN, 2009).

O modelo de covariáveis desenvolvido por Campbell e Mosimann, como explicado por Hijazi e Jernigan (2009), é obtido ao possibilitar que os parâmetros da distribuição estejam em função do vetor de covariável x_i , resultando em uma distribuição Dirichlet condicional de Y_i dado x_i , mutuamente independentes, com parâmetros desconhecidos $\lambda_{ir}(x_i) = (\lambda_{1i}(x_i), \dots, \lambda_{qi}(x_i))$.

A estimação dos parâmetros do modelo, que são os coeficientes de regressão, pode ser feita pelo método de máxima verossimilhança. A função de verossimilhança para uma amostra Y_i , com $i = (1, 2, \dots, n)$ condicionada a covariável x é dada por

$$L(\beta|X, Y) = \prod_{i=1}^n \left\{ \Gamma\left(\sum_{j=1}^q \lambda_{ir}(x_i)\right) \prod_{r=1}^q \frac{y_{ir}^{\lambda_{ir}(x_i)-1}}{\Gamma(\lambda_{ir}(x_i))} \right\} \quad (2.2)$$

Uma parametrização alternativa foi proposta por Maier (2014) para permitir que a distribuição agregasse um parâmetro para média e dispersão. Assim, a heterocedasticidade, ou seja, a variância dos erros, é explicitamente levada em consideração. Nessa parametrização, o parâmetro que indexa a distribuição é modificado para $\lambda_{ir} = \mu_{ir}\phi_i$, onde:

- $\mu_{ir} \in (0, 1)$, representa a média condicional do r -ésimo componente, tal que $\sum_{r=1}^q \mu_{ir} = 1$
- $\phi > 0$ é o parâmetro de precisão que controla a variância condicional das proporções

A função de verossimilhança pode ser representada pela Equação 2.2 com a devida reparametrização, assim dada por

$$f(y_1, \dots, y_q | \mu, \phi) = \prod_{r=1}^q y_r^{\mu_r \phi - 1} \frac{\Gamma\left(\sum_{r=1}^q \mu_r \phi\right)}{\prod_{r=1}^q \Gamma(\mu_r \phi)} \quad (2.3)$$

Com a parametrização definida na Equação 2.3, a composição, portanto, segue uma distribuição de Dirichlet como $Y_i \sim D(\mu_i, \phi_i)$, e o modelo de regressão dado em função desses parâmetros é expresso como

$$g_\mu(\mu) = \eta_\mu = \beta X$$

$$g_\phi(\phi) = \eta_\phi = \gamma Z$$

onde g_μ e g_ϕ são funções de ligação apropriadas para linearização dos parâmetros em função de x_i e z_i .

A variável aleatória X , pode representar somente uma única variável explicativa, mas também pode denotar uma matriz com mais de uma variável explicativa, $X = (X_1, \dots, X_p)$, como apresentado por [Melo, Vasconcellos e Lemonte \(2009\)](#); as componentes Y_i podem ser modeladas por diferentes covariáveis ou ainda, podem se diferenciar em número de covariáveis. Por exemplo, dado X uma matriz de variáveis explicativas contendo três variáveis X_1, X_2 e X_3 , e $Y = (Y_1, Y_2)$, os subconjuntos $X^* = X_1$ e $X^{**} = (X_2, X_3)$ poderiam ser usados para explicar as componentes Y_1 e Y_2 , respectivamente.

A modelagem de dados composicionais em função de uma ou mais variáveis explicativas, portanto, requer a inferência sobre os valores dos parâmetros desconhecidos β_{pr} , associados a cada covariável dos modelos, que são ajustados para cada componente. Uma abordagem clássica para a estimação dos parâmetros da distribuição é a maximização do logaritmo da função de verossimilhança ([WICKER et al., 2008](#)).

Aplicações deste modelo de regressão têm sido desenvolvidas em ampla situações práticas, como o estudo de [Gueorguieva, Rosenheck e Zeltermann \(2008\)](#) que explora a utilização do modelos em dados psiquiátricos, na análise de indicadores econômicos ([LEE; WESTHOFF; THOMPSON, 2023](#); [MORAIS; THOMAS-AGNAN; SIMIONI, 2018](#)) ou no campo da engenharia ([FOROUTAN; AMIRI; KAMRANRAD, 2018](#)).

2.2.1 Estimador de Máxima Verossimilhança

A estimação dos parâmetros por Máxima Verossimilhança consiste em encontrar os valores dos parâmetros que fornecem a maior probabilidade de observar exatamente os dados que foram coletados, ou seja, objetiva escolher parâmetros que maximiza a função de verossimilhança, ou sua forma logarítmica (log-verossimilhança) ([NARAYANAN, 1992](#)).

Dada uma amostra aleatória Y_i com n observações da distribuição de Dirichlet e β o vetor de parâmetros de regressão a ser estimado, a função de verossimilhança total para a distribuição dada pela Equação (2.2) é definida por

$$L(\beta) = \prod_{i=1}^n f(y_i | \lambda_i)$$

Logo, a log-verossimilhança pode ser descrita como

$$\ln L(\beta) = \sum_{i=1}^n \left\{ \ln(\Gamma(\phi_i)) - \sum_{r=1}^q \ln(\Gamma(\lambda_{ir})) + \sum_{r=1}^q \lambda_{ir} \ln(Y_{ir}) \right\} \quad (2.4)$$

Problemas na estimação dos parâmetros por máxima verossimilhança podem ocorrer quando qualquer valor observado na composição Y_r é igual a 0, visto que a presença do termo $\log(Y_{ir})$ em sua função de verossimilhança requer que Y_{ir} seja estritamente positiva para todo i e r . Em vista disso, a fim de evitar a presença de valores zero na variável de interesse, é possível realizar um ajuste nas observações por meio de transformações.

Em sua implementação, [Maier \(2014\)](#) propõe o uso da transformação introduzida por [Smithson e Verkuilen \(2006\)](#) com uma generalização para ajustar as composições quando os valores 0 ou 1 estão presentes. Nessa abordagem os dados são reescalados para distribuírem simetricamente ao redor de 0.5, portanto valores extremos são mais afetados pela transformação, a qual é dada por

$$y_{ir}^* = \frac{y_{ir}(N-1) + q^{-1}}{N}$$

onde y^* é a observação transformada, N é o número de observações e q é o número de componentes.

Os estimadores de máxima verossimilhança (EMV's) podem ser obtidos resolvendo-se o sistema de equações gerados pelas derivadas de primeira ordem da $\ln L(\beta)$ em relação aos parâmetros β :

$$U(\beta) = \frac{\partial \ln(L(\beta))}{\partial \beta_r} = 0$$

onde $U(\beta)$ é denominada função score.

Contudo, o sistema de equações resultante das diferenciações não fornece o resultado direto, ou seja, não há uma expressão de forma fechada para os estimadores de Máxima Verossimilhança, sendo necessário um processo iterativo ou recursivo para a obtenção dos parâmetros ([HIJAZI; JERNIGAN, 2009](#)).

Métodos de otimização numérica são necessários para calcular os estimadores de máxima verossimilhança. O uso dos algoritmos de otimização visa obter uma aproximação iterativa para os valores dos parâmetros que maximizam a função. Os algoritmos de Newton ou quasi-Newton são exemplos de técnicas de otimização não lineares comumente usados na estimação dos parâmetros por máxima verossimilhança ([CORDEIRO; CRIBARI-NETO et al., 2014](#)).

Esses métodos de otimização numérica exigem a escolha de valores iniciais para os parâmetros no início do processo iterativo. A escolha adequada desses valores pode influenciar significativamente a convergência e a precisão da estimação ([HIJAZI; JERNIGAN, 2009](#)).

Segundo [Narayanan \(1992\)](#), uma alternativa prática consiste em utilizar as estimativas obtidas pelo método dos momentos como valores iniciais para os parâmetros da distribuição. Já [Wicker et al. \(2008\)](#) propõe uma aproximação do estimador de máxima verossimilhança como ponto de partida, adaptando-a especificamente para o algoritmo de Newton-Raphson. Por sua vez, [Hijazi e Jernigan \(2009\)](#) sugere o uso de estimativas de mínimos quadrados como valores iniciais.

As propriedades dos estimadores de máxima verossimilhança também se aplicam ao caso do modelo de regressão Dirichlet. Em particular, esses estimadores são assintoticamente normais e assintoticamente não viesados, ou seja, tendem a seguir uma distribuição normal e se aproximar, em média, dos verdadeiros valores dos parâmetros à medida que o tamanho da amostra aumenta ([MELO et al., 2020](#)). Contudo, quando a amostra é pequena, essas boas propriedades nem sempre se mantêm ([CORDEIRO; MCCULLAGH, 1991](#)). Nesses casos, os estimadores podem apresentar viés, ou seja, podem sistematicamente superestimar ou subestimar os verdadeiros parâmetros, e também podem ser ineficientes, isto é, ter variância maior do que a mínima possível. Para contornar esse problema, é possível estimar e corrigir o viés esperado. Essa abordagem será discutida com mais detalhes no Capítulo 3.

3

Correção de viés em modelos de Regressão

Este capítulo aborda a teoria do viés na estimação dos parâmetros da regressão Dirichlet e os principais métodos propostos para sua correção. Inicialmente, é apresentada uma revisão da literatura sobre correção de viés em modelos de regressão. Em seguida, são descritos os métodos investigados neste estudo, a correção analítica de Cox–Snell e as técnicas de reamostragem bootstrap. Por fim, são apresentados os estudos de simulação e a aplicação prática desenvolvidos para avaliar o desempenho dessas metodologias.

3.1 Revisão

A estimação por máxima verossimilhança (EMV) é amplamente utilizada na modelagem estatística devido às suas propriedades assintóticas desejáveis, como consistência, eficiência e normalidade assintótica dos estimadores (MELO; VASCONCELLOS; LEMONTE, 2009). Entretanto, em amostras finitas, especialmente quando o tamanho amostral é pequeno, os EMV's podem apresentar viés não negligenciável (CORDEIRO; CRIBARI-NETO et al., 2014).

O viés é um erro sistemático, observado pela diferença entre o valor esperado do estimador e o verdadeiro valor do parâmetro. Uma das estratégias para reduzir esse erro de estimação, como detalhado por Cordeiro, Cribari-Neto et al. (2014), é a abordagem baseada na dedução da função do viés. Nesse método corretivo, subtrai-se do estimador obtido por máxima verossimilhança uma estimativa do viés calculada com algum grau de aproximação. Por exemplo, usando a função de viés, o estimador corrigido pode ser escrito como:

$$\tilde{\theta} = \hat{\theta} - \hat{B}_1(\hat{\theta})$$

em que θ é o vetor de parâmetros desconhecidos, $\hat{\theta}$ é o EMV de θ e $\hat{B}_1(\hat{\theta})$ denota o viés estimado de segunda ordem. Essa aproximação de segunda ordem do viés do estimador de máxima

verossimilhança pode ser obtida analiticamente, deduzindo uma expressão fechada para o viés, a partir da expressão geral apresentado por [Cox e Snell \(1968\)](#).

Após um mapeamento da literatura, confirma-se que diversos procedimentos têm sido desenvolvidos para reduzir o viés dos estimadores de máxima verossimilhança. Entre eles, destacam-se métodos analíticos, que envolvem expansões de séries de Taylor da função score, ou da função de log-verossimilhança, e métodos numéricos, como o bootstrap. Essas técnicas têm sido amplamente estudadas e adaptadas para diferentes classes de modelos, incluindo regressões lineares e modelos lineares generalizados ([CORDEIRO; CRIBARI-NETO et al., 2014](#), capítulo 4).

O trabalho de [Shen e Yang \(2015\)](#), por exemplo, apresenta uma contribuição sobre correção de viés nos modelos de regressão baseados na distribuição de Weibull, amplamente aplicada em análise de sobrevivência e confiabilidade. Os autores partem do reconhecimento de que o estimador de máxima verossimilhança (EMV) do parâmetro de forma β pode ser fortemente enviesado em situações de censura à direita, especialmente quando há pequeno tamanho amostral, dados censurados ou múltiplas populações Weibull. Nesses contextos, a acurácia da estimação de β torna-se crítica, pois esse parâmetro determina a natureza da distribuição e influencia diretamente as inferências posteriores.

A metodologia proposta pelos autores baseia-se em uma expansão estocástica de terceira ordem aplicada à função de score, permitindo derivar expressões para o viés de segunda e terceira ordem do EMV. A grande vantagem deste enfoque é reduzir um problema multidimensional, com parâmetros de escala e forma, a um problema unidimensional focado na estimativa da forma β , que é a principal fonte de viés. A partir das derivadas da função de score em relação ao parâmetro de interesse, os autores obtêm fórmulas que envolvem expectativas de expressões complexas, cuja avaliação analítica mostra-se inviável.

Para superar esse desafio, o artigo propõe o uso de um bootstrap paramétrico como ferramenta de estimação das quantidades envolvidas na correção de viés. As amostras bootstrap são geradas a partir das distribuições Weibull ajustadas com os EMV's e replicam o mesmo mecanismo de censura presente nos dados originais. A robustez da abordagem é verificada por meio de extensos estudos de simulação de Monte Carlo, que consideram múltiplos cenários de censura, diferentes tamanhos amostrais e números de populações Weibull. Os resultados indicam que tanto a correção de segunda ordem quanto a de terceira ordem conseguem reduzir significativamente o viés do EMV, com destaque para a segunda ordem, que já se mostra suficiente na maioria dos casos. As estimativas corrigidas apresentam melhor desempenho em termos de viés e erro quadrático médio, sendo superiores ao EMV.

O artigo de [Schwartz e Giles \(2016\)](#), por sua vez, apresenta a abordagem analítica para correção de viés do EMV aplicados ao modelo de Poisson inflacionada em zero (ZIP), utilizado na modelagem de dados de contagem com excesso de zeros. Os autores propõem a aplicação da técnica de correção de viés de segunda ordem desenvolvida por Cox e Snell (1968) como uma

alternativa eficaz para aprimorar a qualidade inferencial dos estimadores.

Segundo os autores, o modelo ZIP assume que a variável resposta pode surgir de dois mecanismos: uma distribuição de Bernoulli, e uma distribuição Poisson que modela os valores de contagem (inclusive zeros). A estrutura do modelo resulta em uma função de verossimilhança com duas partes distintas, que são estimadas a partir das derivadas de segunda e terceira ordem da log-verossimilhança.

A eficácia da correção proposta é avaliada por meio de estudos de simulação de Monte Carlo, abrangendo diversos cenários de tamanhos amostrais (50, 60, 75, 100, 125, 150 e 200) e combinações dos parâmetros do modelo (π e λ). Os resultados mostram que os estimadores corrigidos, seja analiticamente seja por bootstrap, apresentam viés reduzido. Também observou-se uma redução no EQM. Além disso, a abordagem analítica dominou a abordagem bootstrap, exceto nas situações em que os estimadores são mais enviesados.

Mazucheli, Menezes e Chakraborty (2019) introduz uma nova proposta de modelagem para dados proporcionais, a partir da distribuição unit-Lindley com um único parâmetro, que assume valores em (0,1). O foco do artigo é o desenvolvimento teórico da distribuição e a aplicação de um modelo de regressão com resposta contínua limitada ao intervalo unitário, tendo como base essa nova distribuição.

No contexto da regressão, o parâmetro da distribuição unit-Lindley, denotado por θ , é relacionado a covariáveis por meio de uma função de ligação logarítmica. A modelagem é feita por meio da máxima verossimilhança. No entanto, eles também destacam que o estimador de máxima verossimilhança (EMV) pode ser enviesado em amostras pequenas, o que motiva a incorporação de técnicas de correção de viés no processo inferencial.

Para esse fim, os autores aplicam a metodologia de correção de viés analítica de Cox e Snell (1968), obtendo-se uma expressão explícita para o viés de segunda ordem do EMV. A fórmula geral do viés é apresentada, envolvendo derivadas de primeira, segunda e terceira ordens da função de log-verossimilhança, bem como elementos da matriz de informação de Fisher. O viés é então subtraído do EMV, resultando em um estimador corrigido com menor viés em amostras pequenas.

Os autores realizam uma análise de simulação de Monte Carlo, comparando o EMV e o estimador corrigido por Cox e Snell. As métricas utilizadas incluem viés médio, erro quadrático médio (EQM) e cobertura dos intervalos de confiança. Os resultados indicam que o viés do EMV é sistematicamente positivo, e que o estimador corrigido apresenta viés reduzido e melhor desempenho em termos de EQM.

No mesmo contexto da correção de viés pelo método analítico, o artigo de Godwin e Giles (2019), com foco em situações em que a função de viés não é constante em relação ao parâmetro populacional, Palm, Bayer e Cintra (2020) aplicado ao modelo de regressão Rayleigh introduzido para modelar a amplitude de pixels em imagens, e Magalhães, Gallardo e Bourguignon (2021)

com ênfase nos parâmetros do modelo de Regressão Gama Inversa (RIG), são exemplos recentes da aplicabilidade desse método de correção.

Nos três cenários os autores baseiam-se no método de Cox e Snell para dedução da expressão de viés ou refinamento da abordagem. [Godwin e Giles \(2019\)](#), por exemplo, apresentam como contribuição o desenvolvimento de uma abordagem alternativa, que considera explicitamente a variabilidade da função de viés. Os autores mostram que as expansões da séries de Taylor e cumulantes da log-verossimilhança, podem levar a resultados enganosos quando a função de viés muda rapidamente ou possui curvatura acentuada. Para contornar isso, eles propõem um ajuste à expressão padrão, aumentando a acurácia das correções.

O novo método é aplicado em dois contextos principais: (i) a distribuição exponencial com parâmetro inverso, e (ii) a regressão com erro de medição. Em ambos os casos, os autores derivam as expressões do viés analítico ajustado, com base na manipulação algébrica explícita. Além da dedução teórica, os autores realizam estudos de simulação de Monte Carlo com diferentes tamanhos amostrais e configurações paramétricas, a fim de comparar o desempenho do método proposto com o de Cox e Snell. Os resultados sugerem que a nova abordagem oferece reduções consideravelmente maiores no viés e melhor estabilidade, especialmente em casos com funções de viés fortemente não lineares. O ponto de partida da metodologia é a constatação de que, em muitos modelos estatísticos aplicados, a estrutura do viés pode mudar acentuadamente em relação ao parâmetro, e isso não é bem tratado pelas fórmulas clássicas como as de [Cox e Snell \(1968\)](#).

Já o trabalho de [Palm, Bayer e Cintra \(2020\)](#) compara a correção analítica de Cox e Snell com o método de modificação da função score de Firth, e o Bootstrap paramétrico. As simulações e experimentos com dados reais mostram que os métodos corrigidos, especialmente o de Firth, melhoram a qualidade das estimativas em cenários com poucos dados.

Outrossim, o artigo de [Magalhães, Gallardo e Bourguignon \(2021\)](#) centra-se na melhoria da estimação dos parâmetros de regressão e da dispersão em situações em que a variável resposta representa tempos de sobrevivência, custos ou velocidades, modeladas pela distribuição assimétrica gama. As extensivas simulações de Monte Carlo conduzidas pelos autores demonstram a redução de viés na maioria dos cenários comparados, sendo recomendado pelos autores o uso especificamente da correção analítica de Cox e Snell ou a correção de Firth para melhorar a estimação dos parâmetros do modelo RIG.

O método de correção numérica também está presente na literatura recente. Além da aplicabilidade do Bootstrap, a técnica de Jackknife foi apresentada nos estudos de [Türkan e Özel \(2018\)](#) e [Seifollahi, Bevrani e Albalawi \(2024\)](#). A abordagem consiste em aplicar o método numérico de jackknife para correção de primeira ordem. Essa técnica envolve a remoção sistemática de uma observação da amostra, recalculando os estimadores EMV em cada amostra reduzida. A média dessas estimativas parciais é então utilizada para construir um estimador com viés reduzido ([SEIFOLLAHI; BEVRANI; ALBALAWI, 2024](#)).

Türkan e Özel (2018) aplicam essa técnica para corrigir o viés dos parâmetros do modelo de regressão ridge quando a variável resposta possui distribuição binomial negativa, propondo assim uma modificação da regressão originária. Para isso, estimativas EMV são obtidas por maximização numérica da log-verossimilhança, o que envolve rotinas de otimização computacional padrão. Para avaliar o desempenho do estimador jackknife, os autores conduzem uma simulação de Monte Carlo com diferentes tamanhos amostrais e valores do parâmetro do modelo de regressão. Os autores concluem que há um efetivo ganho na redução do viés com o procedimento de jackknife.

Por fim, o trabalho de Seifollahi, Bevrani e Albalawi (2024) aplica a estratégia de correção de viés por meio do método Jackknife ao estimador *Liu-type* para os parâmetros do modelo de regressão beta. Esse estimador é uma alternativa utilizada para mitigar o problema de multicolinearidade; entretanto, assim como o estimador de máxima verossimilhança, o estimador *Liu-type* pode apresentar viés. Com o objetivo de eliminar o viés de primeira ordem induzido por esse estimador, os autores propõem uma versão modificada pelo método Jackknife, a qual apresentou melhor desempenho em termos de erro quadrático médio quando comparada a variações baseadas nos estimadores Liu, *Liu-type* e Ridge.

3.2 Correção do viés do Modelo de Regressão de Dirichlet

Assim como ocorre em outros contextos, os estimadores obtidos por máxima verossimilhança no modelo de Regressão Dirichlet também podem ser tendenciosos em amostras de tamanho reduzido. A necessidade de corrigir o viés nesse cenário motiva o desenvolvimento de métodos específicos para modelos Dirichlet, adaptando técnicas existentes de correção de viés.

Embora a literatura sobre correção de viés em modelos de regressão seja extensa, os estudos específicos em modelos de regressão Dirichlet ainda são relativamente escassos. A pesquisa recente de Melo et al. (2020) tem abordado essa lacuna, introduzindo um procedimento analítico para correção e redução do viés na estimação dos parâmetros da regressão Dirichlet. A metodologia apresentada pelos referidos autores é discutida na seção 3.2.2.

3.2.1 Viés do Estimador de Máxima Verossimilhança

Considere um modelo estatístico com um vetor de parâmetros desconhecidos $\theta = (\theta_1, \dots, \theta_t)$. Se denotar $\hat{\theta}$ o estimador de máxima verossimilhança desses parâmetros, o viés do estimador é definido como a diferença entre o valor esperado da estimativa e o valor verdadeiro do parâmetro, que pode ser escrito como apresenta Ferrari e Cribari-Neto (1998):

$$B(\hat{\theta}) = E(\hat{\theta} - \theta)$$

Esse viés pode ser expresso como uma série em função do tamanho da amostra n , conforme apresentado por (FERRARI; CRIBARI-NETO, 1998):

$$B(\hat{\theta}) = \frac{B_1(\theta)}{n} + \frac{B_2(\theta)}{n^2} + \frac{B_3(\theta)}{n^3} + \dots$$

Os termos $B_1(\theta)$, $B_2(\theta)$ e $B_3(\theta)$ são obtidos a partir de expansões de Taylor da função escore $U(\theta)$ (definida anteriormente na Seção 2.2.1). O primeiro termo da série $B_1(\theta)/n$ é dito ser o viés de segunda ordem para $\hat{\theta}$, e costuma ser o principal alvo das técnicas de correção de viés em modelos com amostras finitas.

A expressão geral para o viés de segunda ordem dos estimadores de máxima verossimilhança em modelos com múltiplos parâmetros deduzida por Cox e Snell (1968) tornou-se uma referência na literatura (ver seção 3.1) e tem sido utilizada como base para correções de viés em diversas classes de modelos. Por exemplo, Vasconcellos e Dourado (2008) derivam a expressão para correção de viés nos estimadores dos parâmetros da família exponencial biparamétrica, Simas, Barreto-Souza e Rocha (2010) deduzem a função para a classe de modelos de regressão beta, Menezes et al. (2021) adaptam a equação para a distribuição Unit-Weibull.

A expressão fechada para a estimação do viés de segunda ordem baseada em Cox e Snell (1968), quando a amostra é independente, mas não necessariamente identicamente distribuída, pode ser escrita conforme a notação apresentada por Cordeiro e Klein (1994):

$$B_1(\hat{\theta}_u) = \sum_{j,k,l} \kappa^{uj} \kappa^{kl} \left\{ \kappa_{jk}^{(l)} - \frac{1}{2} k_{jkl} \right\}, \quad u = 1, \dots, t \quad (3.1)$$

onde:

- ℓ representa a função de log-verossimilhança;
- $\hat{\theta}_u$ é o estimador de máxima verossimilhança para o u -ésimo termo de $\theta = (\theta_1, \dots, \theta_t)$;
- $\kappa_{jk} = E \left(\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} \right)$ é o valor esperado da segunda derivada da log-verossimilhança;
- $\kappa_{jkl} = \left(\frac{\partial^3 \ell}{\partial \theta_j \partial \theta_k \partial \theta_l} \right)$ é a terceira derivada esperada da log-verossimilhança;
- $k_{jk}^{(l)} = \frac{\partial}{\partial \theta_l} \kappa_{jk}$ é a derivada de κ_{jk} em relação a θ_l
- $K = -\kappa_{jk}$ é a matriz de informação de Fisher e $K^{-1} = \{-\kappa^{jk}\}$ é a sua inversa.

Portanto, o viés é obtido como solução de um sistema envolvendo a inversa da matriz de informação de Fisher e os termos de derivadas da função de verossimilhança.

Baseado nos resultados obtido para o viés de segunda ordem, $B_1(\hat{\theta})$, o estimador corrigido para $\hat{\theta}$, como apresentado por Cordeiro e McCullagh (1991), é definido como $\hat{\theta}_{BC} = \hat{\theta} - \hat{B}_1(\theta)$, onde $\hat{B}_1(\hat{\theta})$ é o valor estimado de $B_1(\hat{\theta})$ avaliado em $\hat{\theta}$.

3.2.2 Correção Analítica

Usando abordagem analítica, [Melo et al. \(2020\)](#) propuseram uma expressão fechada para o viés de segunda ordem do EMV do modelo de regressão multivariado de Dirichlet, baseado na fórmula geral de Cox e Snell ([1968](#)) para correção do estimador.

Considerando a formulação dos autores, e sendo Y_{iq} um vetor q -dimensional com distribuição Dirichlet, definida conforme seção 2.2, e $\beta = \text{vec}(\beta_{i1}, \dots, \beta_{iq})$ o vetor de coeficientes de regressão desconhecidos q -dimensional, com $\text{vec}(\cdot)$ sendo um operador que transforma a matriz de parâmetros em um vetor, empilhando as colunas da matriz, a expressão do viés de segunda ordem deduzida pelos autores é, em notação matricial, dada por

$$B(\hat{\beta}) = [(I_q \otimes X)^T W (I_q \otimes X)]^{-1} M \text{vec} \left([(I_q \otimes X)^T W (I_q \otimes X)]^{-1} \right) \quad (3.2)$$

onde:

- I_q é a matriz identidade $q \times q$;
- X é a matriz de variáveis independentes do modelo;
- W é uma matriz $nq \times nq$ definida na forma particionada, onde cada elemento $a = (1, \dots, q)$ e $b = (1, \dots, q)$ de W é uma matriz diagonal com o i -ésimo elemento dado por

$$\begin{cases} -g'(\eta_{ia})^2 [\psi(\phi_r) - \phi'(\lambda_{ia})], & a = b \\ -g'(\eta_{ia})g'(\eta_{ib})\phi'(\phi_r), & a \neq b \end{cases}$$

- e, o símbolo $\otimes$ representa o produto de Kronecker.

Contudo, as definições apresentadas por [Melo et al. \(2020\)](#) carecem de importantes etapas para a correta interpretação e replicação direta de sua expressão. Conduzindo a resultados incoerentes na implementação das simulação. Diante disso, com o objetivo de possibilitar a análise comparativa dos métodos, o presente estudo apresenta uma formulação analítica também fundamentada no arcabouço teórico de Cox-Snell, como segue.

Inicialmente, é adotada uma reindexação na representação vetorial dos coeficientes de regressão, a fim de evitar índices duplos:

$$u, j, k, l = (r - 1)p + a, \quad r \in \{1, \dots, q\}, \quad a \in \{1, \dots, p\}.$$

Assim, o vetor de coeficientes de regressão, que tem pq componentes, pode ser reescrito como

$$\beta^T = (\beta_1, \beta_2, \dots, \beta_{pq}),$$

onde pq é o produto de p e q , que é o número de elementos do vetor β .

Considerando a reindexação, a fórmula apresentada na Equação (3.1) pode ser definida para o vetor de parâmetros da regressão Dirichlet com $u \in \{1, \dots, pq\}$. O viés pode ser escrito em notação matricial, como define [Cordeiro e Klein \(1994\)](#):

$$B(\beta) = K^{-1} A \text{Vec} \left(K^{-1} \right), \quad (3.3)$$

Onde $A = [A^{(1)} | A^{(2)} | \dots | A^{(pq)}]$, $A^{(l)}$ é a matriz cujo elemento (j, k) é

$$a_{jk}^{(l)} = \kappa_{jk}^{(l)} - \frac{1}{2} \kappa_{jkl} \quad (3.4)$$

A função score é dada pela vetor de derivadas parciais da log-verossimilhança, dado por

$$U(\beta)^T = \left(\frac{\partial l}{\partial \beta_{11}}, \frac{\partial l}{\partial \beta_{21}}, \dots, \frac{\partial l}{\partial \beta_{p1}}, \dots, \frac{\partial l}{\partial \beta_{1q}}, \frac{\partial l}{\partial \beta_{2q}}, \dots, \frac{\partial l}{\partial \beta_{pq}} \right), \quad (3.5)$$

onde

$$\begin{aligned} \frac{\partial l}{\partial \beta_{ar}} &= \sum_{i=1}^n \frac{\partial \ln(\Gamma(\phi_i))}{\partial \phi_i} \frac{\partial \phi_i}{\partial \lambda_{ir}} \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial \eta_{ir}}{\partial \beta_{ar}} - \sum_{i=1}^n \frac{\partial \ln(\Gamma(\lambda_{ir}))}{\partial \lambda_{ir}} \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial \eta_{ir}}{\partial \beta_{ar}} \\ &+ \sum_{i=1}^n \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial \eta_{ir}}{\partial \beta_{ar}} \ln(Y_{ir}) = \\ &= \sum_{i=1}^n \psi(\phi_i) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} x_{ia} - \sum_{i=1}^n \psi(\lambda_{ir}) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} x_{ia} + \sum_{i=1}^n \ln(Y_{ir}) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} x_{ia} \\ &= \sum_{i=1}^n (\psi(\phi_i) - \psi(\lambda_{ir}) + \ln(Y_{ir})) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} x_{ia}. \end{aligned} \quad (3.6)$$

A matriz de Informação de Fisher observada pode ser escrita como

$$J(\beta) = - \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_{11}^2} & \frac{\partial^2 l}{\partial \beta_{21} \partial \beta_{11}} & \dots & \frac{\partial^2 l}{\partial \beta_{p1} \partial \beta_{11}} & \dots & \frac{\partial^2 l}{\partial \beta_{1q} \partial \beta_{11}} & \frac{\partial^2 l}{\partial \beta_{2q} \partial \beta_{11}} & \dots & \frac{\partial^2 l}{\partial \beta_{pq} \partial \beta_{11}} \\ \frac{\partial^2 l}{\partial \beta_{11} \partial \beta_{21}} & \frac{\partial^2 l}{\partial \beta_{21}^2} & \dots & \frac{\partial^2 l}{\partial \beta_{p1} \partial \beta_{21}} & \dots & \frac{\partial^2 l}{\partial \beta_{1q} \partial \beta_{21}} & \frac{\partial^2 l}{\partial \beta_{2q} \partial \beta_{21}} & \dots & \frac{\partial^2 l}{\partial \beta_{pq} \partial \beta_{21}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 l}{\partial \beta_{11} \partial \beta_{p1}} & \frac{\partial^2 l}{\partial \beta_{21} \partial \beta_{p1}} & \dots & \frac{\partial^2 l}{\partial \beta_{p1}^2} & \dots & \frac{\partial^2 l}{\partial \beta_{1q} \partial \beta_{p1}} & \frac{\partial^2 l}{\partial \beta_{2q} \partial \beta_{p1}} & \dots & \frac{\partial^2 l}{\partial \beta_{pq} \partial \beta_{p1}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 l}{\partial \beta_{11} \partial \beta_{1q}} & \frac{\partial^2 l}{\partial \beta_{21} \partial \beta_{1q}} & \dots & \frac{\partial^2 l}{\partial \beta_{p1} \partial \beta_{1q}} & \dots & \frac{\partial^2 l}{\partial \beta_{1q}^2} & \frac{\partial^2 l}{\partial \beta_{2q} \partial \beta_{1q}} & \dots & \frac{\partial^2 l}{\partial \beta_{pq} \partial \beta_{1q}} \\ \frac{\partial^2 l}{\partial \beta_{11} \partial \beta_{2q}} & \frac{\partial^2 l}{\partial \beta_{21} \partial \beta_{2q}} & \dots & \frac{\partial^2 l}{\partial \beta_{p1} \partial \beta_{2q}} & \dots & \frac{\partial^2 l}{\partial \beta_{1q} \partial \beta_{2q}} & \frac{\partial^2 l}{\partial \beta_{2q}^2} & \dots & \frac{\partial^2 l}{\partial \beta_{pq} \partial \beta_{2q}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \beta_{11} \partial \beta_{pq}} & \frac{\partial^2 l}{\partial \beta_{21} \partial \beta_{pq}} & \dots & \frac{\partial^2 l}{\partial \beta_{p1} \partial \beta_{pq}} & \dots & \frac{\partial^2 l}{\partial \beta_{1q} \partial \beta_{pq}} & \frac{\partial^2 l}{\partial \beta_{2q} \partial \beta_{pq}} & \dots & \frac{\partial^2 l}{\partial \beta_{pq}^2} \end{pmatrix}, \quad (3.7)$$

onde o elemento $-J_{jk}$, para $j = (r-1)p + a$ e $k = (s-1)p + b$ e $a, b = (1, \dots, p)$ e $s = (1, \dots, q)$ é dado por

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_{br} \partial \beta_{ar}} &= \sum_{i=1}^n \psi'(\phi_i) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 x_{ia} x_{ib} + \sum_{i=1}^n \psi(\phi_i) \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} - \\
&- \sum_{i=1}^n \psi'(\lambda_{ir}) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 x_{ia} x_{ib} - \sum_{i=1}^n \psi(\lambda_{ir}) \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} + \\
&+ \sum_{i=1}^n \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} \ln(Y_{ir}) \\
&= \sum_{i=1}^n (\psi'(\phi_i) - \psi'(\lambda_{ir})) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 x_{ia} x_{ib} + \\
&+ \sum_{i=1}^n (\psi(\phi_i) - \psi(\lambda_{ir}) + \ln(Y_{ir})) \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib}, \quad r = s. \tag{3.8}
\end{aligned}$$

e

$$\frac{\partial^2 l}{\partial \beta_{bs} \partial \beta_{ar}} = \sum_{i=1}^n \psi'(\phi_i) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial \lambda_{is}}{\partial \eta_{is}} x_{ia} x_{ib}, \quad r \neq s. \tag{3.9}$$

A matriz de Informação de Fisher esperada é a matriz $K(\beta) = E[J(\beta)]$, com

$$\begin{aligned}
\kappa_{brar} = E \left[\frac{\partial^2 l}{\partial \beta_{br} \partial \beta_{ar}} \right] &= \sum_{i=1}^n \psi'(\phi_i) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 x_{ia} x_{ib} - \sum_{i=1}^n \psi'(\lambda_{ir}) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 x_{ia} x_{ib} \\
&= \sum_{i=1}^n (\psi'(\phi_i) - \psi'(\lambda_{ir})) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 x_{ia} x_{ib}, \quad r = s \tag{3.10}
\end{aligned}$$

Dado $E[\ln(Y_{ir})] = \psi(\lambda_{ir}) - \psi(\phi_i)$, então

$$\kappa_{bsar} = E \left[\frac{\partial^2 l}{\partial \beta_{bs} \partial \beta_{ar}} \right] = \frac{\partial^2 l}{\partial \beta_{bs} \partial \beta_{ar}} = \sum_{i=1}^n \psi'(\phi_i) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial \lambda_{is}}{\partial \eta_{is}} x_{ia} x_{ib}, \quad r \neq s. \tag{3.11}$$

A matriz de derivadas de terceira ordem e valores esperados é dada pelas equações:

$$\begin{aligned}
\frac{\partial^3 l}{\partial \beta_{cr} \partial \beta_{br} \partial \beta_{ar}} &= \sum_{i=1}^n \psi''(\phi_i) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^3 x_{ia} x_{ib} x_{ic} + 3 \sum_{i=1}^n \psi'(\phi_i) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} x_{ic} + \\
&+ \sum_{i=1}^n \psi(\phi_i) \frac{\partial^3 \lambda_{ir}}{\partial \eta_{ir}^3} x_{ia} x_{ib} x_{ic} - \sum_{i=1}^n \psi''(\lambda_{ir}) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^3 x_{ia} x_{ib} x_{ic} - \\
&- 3 \sum_{i=1}^n \psi'(\lambda_{ir}) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} x_{ic} - \sum_{i=1}^n \psi(\lambda_{ir}) \frac{\partial^3 \lambda_{ir}}{\partial \eta_{ir}^3} x_{ia} x_{ib} x_{ic} + \\
&+ \sum_{i=1}^n \frac{\partial^3 \lambda_{ir}}{\partial \eta_{ir}^3} \ln(Y_{ir}) x_{ia} x_{ib} x_{ic} \\
&= \sum_{i=1}^n (\psi''(\phi_i) - \psi''(\lambda_{ir})) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^3 x_{ia} x_{ib} x_{ic} + \\
&+ 3 \sum_{i=1}^n (\psi'(\phi_i) - \psi'(\lambda_{ir})) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} x_{ic} + \\
&+ \sum_{i=1}^n (\psi(\phi_i) - \psi(\lambda_{ir}) + \ln(Y_{ir})) \frac{\partial^3 \lambda_{ir}}{\partial \eta_{ir}^3} x_{ia} x_{ib} x_{ic}. \tag{3.12}
\end{aligned}$$

$$\begin{aligned}
\kappa_{crbrar} = E \left[\frac{\partial^3 l}{\partial \beta_{cr} \partial \beta_{br} \partial \beta_{ar}} \right] &= \sum_{i=1}^n (\psi''(\phi_i) - \psi''(\lambda_{ir})) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^3 x_{ia} x_{ib} x_{ic} + \\
&+ 3 \sum_{i=1}^n (\psi'(\phi_i) - \psi'(\lambda_{ir})) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} x_{ic}, \tag{3.13} \\
r = s \text{ e } r = t.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^3 l}{\partial \beta_{ct} \partial \beta_{br} \partial \beta_{ar}} &= \sum_{i=1}^n \psi''(\phi_i) \frac{\partial \lambda_{it}}{\partial \eta_{it}} \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 x_{ia} x_{ib} x_{ic} + \\
&+ \sum_{i=1}^n \psi'(\phi_i) \frac{\partial \lambda_{it}}{\partial \eta_{it}} \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} x_{ic}, \quad r \neq t. \tag{3.14}
\end{aligned}$$

$$\kappa_{ctbrar} = E \left[\frac{\partial^3 l}{\partial \beta_{ct} \partial \beta_{br} \partial \beta_{ar}} \right] = \frac{\partial^3 l}{\partial \beta_{ct} \partial \beta_{br} \partial \beta_{ar}}, \quad r \neq t. \tag{3.15}$$

$$\frac{\partial^3 l}{\partial \beta_{ct} \partial \beta_{bs} \partial \beta_{ar}} = \sum_{i=1}^n \psi''(\phi_i) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial \lambda_{is}}{\partial \eta_{is}} \frac{\partial \lambda_{it}}{\partial \eta_{it}} x_{ia} x_{ib} x_{ic}, \quad r \neq s, r \neq t \text{ e } s \neq t. \tag{3.16}$$

$$\kappa_{ctbsar} = E \left[\frac{\partial^3 l}{\partial \beta_{ct} \partial \beta_{bs} \partial \beta_{ar}} \right] = \frac{\partial^3 l}{\partial \beta_{ct} \partial \beta_{bs} \partial \beta_{ar}}, \quad r \neq s, r \neq t \text{ e } s \neq t. \tag{3.17}$$

Os elementos da matriz de cumulantes da log-verossimilhança são dados por

$$\begin{aligned} \kappa_{brar}^{(cr)} &= \frac{\partial \kappa_{brar}}{\partial \beta_{cr}} = \sum_{i=1}^n (\psi''(\phi_i) - \psi''(\lambda_{ir})) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^3 x_{ia} x_{ib} x_{ic} + \\ &+ 2 \sum_{i=1}^n (\psi'(\phi_i) - \psi'(\lambda_{ir})) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} x_{ia} x_{ib} x_{ic}. \end{aligned} \quad (3.18)$$

$$\kappa_{brar}^{(cs)} = \frac{\partial \kappa_{brar}}{\partial \beta_{cs}} = \sum_{i=1}^n \psi''(\phi_i) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 \frac{\partial \lambda_{is}}{\partial \eta_{is}} x_{ia} x_{ib} x_{ic}. \quad (3.19)$$

$$\begin{aligned} \kappa_{bsar}^{(cr)} &= \frac{\partial \kappa_{bsar}}{\partial \beta_{cr}} = \sum_{i=1}^n \psi''(\phi_i) \left(\frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \right)^2 \frac{\partial \lambda_{is}}{\partial \eta_{is}} x_{ia} x_{ib} x_{ic} + \\ &+ \sum_{i=1}^n \psi'(\phi_i) \frac{\partial^2 \lambda_{ir}}{\partial \eta_{ir}^2} \frac{\partial \lambda_{is}}{\partial \eta_{is}} x_{ia} x_{ib} x_{ic} \end{aligned} \quad (3.20)$$

$$\kappa_{bsar}^{(ct)} = \frac{\partial \kappa_{bsar}}{\partial \beta_{ct}} = \sum_{i=1}^n \psi''(\phi_i) \frac{\partial \lambda_{ir}}{\partial \eta_{ir}} \frac{\partial \lambda_{is}}{\partial \eta_{is}} \frac{\partial \lambda_{is}}{\partial \eta_{is}} x_{ia} x_{ib} x_{ic}. \quad (3.21)$$

3.2.2.1 Simulação Monte Carlo

Com a finalidade de validar a formulação proposta e comparar com resultados reportados por [Melo et al. \(2020\)](#), um modelo especificado pela Equação (3.22) foi avaliado usando o método de Simulação Monte Carlo para criação de amostras aleatórias, com 10.000 réplicas. As variáveis preditoras x_{ip} foram geradas aleatoriamente segundo uma distribuição Uniforme Padrão $U(0, 1)$

$$\log(\lambda_{ir}) = \eta_{ir} = \beta_{1r} x_{i1}, \dots, \beta_{pr} x_{ip} \quad (3.22)$$

O modelo foi especificado considerando $p = q = 3$, e os resultados numéricos encontram-se apresentados na Tabela 1. De modo geral, observa-se que os estimadores corrigidos apresentam menor viés quando comparados aos estimadores obtidos pelo método de Máxima Verossimilhança. Por exemplo, para o tamanho amostral igual a 20, o viés relativo estimado para o parâmetro β_{22} foi de 0.174, enquanto, após a correção, o viés foi oito vezes menor em relação ao estimador de Máxima Verossimilhança (0.021). Esse resultado evidencia a efetividade do método proposto na estimação dos parâmetros, indicando que os estimadores corrigidos se encontram mais próximos dos valores verdadeiros quando comparados aos estimadores obtidos pelo EMV.

Adicionalmente, a precisão das estimativas, avaliada por meio do erro quadrático médio (EQM), também apresentou melhora significativa para os tamanhos amostrais reduzidos ($n=20$ e

Tabela 1 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Cox-Snell para o modelo linearmente especificado e sem intercepto, $p = q = 3$.

Parameter	$N = 20$						$N = 30$						$N = 40$						$N = 50$					
	B			B_{cs}			B			B_{cs}			B			B_{cs}			B			B_{cs}		
	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}
β_{11}	0.195	0.041	1.143	1.024	0.113	0.013	0.726	0.674	0.077	0.002	0.587	0.555	0.059	0.000	0.492	0.471			0.059	0.000	0.492	0.471		
β_{21}	0.175	0.022	1.115	0.999	0.100	0.000	0.714	0.665	0.094	0.018	0.587	0.554	0.066	0.006	0.494	0.472			0.066	0.006	0.494	0.472		
β_{31}	0.205	0.052	1.137	1.017	0.126	0.024	0.735	0.680	0.074	-0.001	0.580	0.550	0.071	0.010	0.498	0.476			0.071	0.010	0.498	0.476		
β_{12}	0.204	0.050	1.149	1.028	0.115	0.015	0.728	0.676	0.079	0.004	0.582	0.551	0.062	0.003	0.490	0.469			0.062	0.003	0.490	0.469		
β_{22}	0.174	0.021	1.125	1.010	0.097	-0.003	0.714	0.665	0.091	0.015	0.588	0.556	0.064	0.004	0.493	0.471			0.064	0.004	0.493	0.471		
β_{32}	0.199	0.046	1.136	1.017	0.123	0.021	0.732	0.677	0.074	0.000	0.572	0.542	0.068	0.007	0.495	0.473			0.068	0.007	0.495	0.473		
β_{13}	0.192	0.038	1.133	1.015	0.107	0.007	0.727	0.676	0.073	-0.002	0.582	0.552	0.060	0.001	0.492	0.471			0.060	0.001	0.492	0.471		
β_{23}	0.178	0.025	1.116	1.001	0.101	0.001	0.723	0.673	0.092	0.016	0.592	0.559	0.068	0.008	0.489	0.467			0.068	0.008	0.489	0.467		
β_{33}	0.204	0.051	1.131	1.013	0.130	0.028	0.734	0.678	0.077	0.003	0.580	0.549	0.066	0.006	0.493	0.472			0.066	0.006	0.493	0.472		
Parameter	$N = 60$						$N = 100$						$N = 500$						$N = 1000$					
	B			B_{cs}			B			B_{cs}			B			B_{cs}			B			B_{cs}		
	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}	B	EQM	EQM _{cs}
β_{11}	0.059	0.009	0.439	0.422	0.032	0.002	0.319	0.312	0.006	0.000	0.135	0.135	0.005	0.002	0.093	0.093			0.005	0.002	0.093	0.093		
β_{21}	0.053	0.009	0.437	0.422	0.028	-0.001	0.318	0.311	0.005	-0.001	0.134	0.134	0.002	-0.001	0.094	0.094			0.002	-0.001	0.094	0.094		
β_{31}	0.042	0.006	0.439	0.424	0.032	0.002	0.317	0.310	0.005	-0.001	0.133	0.133	0.002	-0.001	0.094	0.094			0.002	-0.001	0.094	0.094		
β_{12}	0.059	0.003	0.439	0.423	0.035	0.006	0.320	0.313	0.006	0.000	0.135	0.134	0.004	0.001	0.093	0.093			0.004	0.001	0.093	0.093		
β_{22}	0.053	0.003	0.438	0.423	0.028	-0.002	0.321	0.314	0.007	0.001	0.134	0.133	0.002	-0.001	0.094	0.094			0.002	-0.001	0.094	0.094		
β_{32}	0.043	0.005	0.437	0.422	0.029	0.000	0.317	0.310	0.005	-0.001	0.133	0.132	0.002	-0.001	0.094	0.094			0.002	-0.001	0.094	0.094		
β_{13}	0.056	-0.008	0.440	0.424	0.033	0.003	0.318	0.311	0.007	0.001	0.134	0.134	0.005	0.002	0.094	0.093			0.005	0.002	0.094	0.093		
β_{23}	0.055	-0.006	0.432	0.416	0.028	-0.002	0.319	0.312	0.006	0.000	0.134	0.133	0.001	0.001	0.094	0.093			0.001	0.001	0.094	0.093		
β_{33}	0.046	-0.004	0.437	0.422	0.030	0.001	0.316	0.309	0.004	-0.002	0.133	0.133	0.002	0.001	0.094	0.093			0.002	0.001	0.094	0.093		

Fonte: Produzido pelos autores.

30). Nesse contexto, os estimadores corrigidos apresentaram redução do erro quadrático médio, o qual diminuiu gradativamente com o aumento do tamanho amostral, conforme esperado na literatura revisada e também observado no estudo de [Melo et al. \(2020\)](#). Observou-se ainda que, para tamanhos amostrais maiores, os valores de viés e EQM entre os estimadores com e sem o método de correção tornam-se aproximadamente equivalentes, confirmando a equivalência prática das abordagens no contexto do modelo de regressão Dirichlet quando o tamanho amostral é suficientemente grande. Tal conclusão se assemelha às apresentadas no estudo de [Melo et al. \(2020\)](#), o que demonstra a efetividade das definições apresentadas nesse estudo, as quais serão posteriormente comparados ao método de reamostragem de bootstrap.

3.2.3 Correção por Bootstrap

O método bootstrap descrito no estudo seminal de [Efron \(1979\)](#) foi introduzido como uma extensão e aprimoramento conceitual do método de reamostragem de jackknife (descrito em [Miller \(1974\)](#)) apresentando-se como uma alternativa mais aplicável, e confiável para a avaliação da variabilidade estatística. Efron (1979) fundamenta o bootstrap como uma abordagem baseada em simulação computacional para a atribuição de medidas de precisão a estimativas estatísticas. O procedimento consiste em obter B amostras independentes $(Y_1^*, Y_2^*, \dots, Y_B^*)$, chamadas amostras de bootstrap, com tamanho n de observações reamostradas, com reposição, a partir do conjunto de dados originalmente observado Y , e assim obtendo-se, para cada amostra bootstrap, a estatística de interesse.

Essa técnica de reamostragem mostra-se como uma alternativa para obtenção de aproximações empíricas confiáveis, sem recorrer a realização de extensos cálculos analíticos, que muitas vezes baseiam-se em suposições teóricas difíceis de verificar, especialmente em contextos com amostras pequenas ou modelos complexos, onde soluções teóricas podem ser inacessíveis. A validade das aproximações obtidas pelo bootstrap é justificada pela Lei dos Grandes Números, que garante que as estatísticas derivadas das reamostragens convergem para os valores verdadeiros sob o modelo ajustado à medida que o número de reamostragem aumenta ([DAVISON; HINKLEY, 1997](#)).

O bootstrap pode ser aplicado em contextos não paramétricos, como o apresentado por [Efron \(1979\)](#), quando não se assume um modelo formal para a distribuição da variável resposta, mas também em contextos paramétricos, quando há um modelo probabilístico bem especificado, extensão apresentada por [Efron e Tibshirani \(1994\)](#). Em ambos os casos, aplicações do método mostram sua usabilidade para estimar propriedades estatística, como viés, variância, distribuição de estimadores, intervalos de confiança e testes de hipóteses.

Na correção de viés dos estimadores de máxima verossimilhança, [Ferrari e Cribari-Neto \(1998\)](#), ao explorar a relação entre o bootstrap e as correções analíticas por meio de expansões em séries, mostram que o bootstrap fornece uma forma natural e prática de definir e aplicar correções de viés em amostras finitas. O estudo oferece tanto uma formulação multiplicativa da correção de

viés, quanto uma expressão explícita para o viés de terceira ordem dos estimadores em modelos multiparamétricos. Além disso, demonstram que o bootstrap iterado, onde o processo é repetido diversas vezes, conduz a aproximações de ordens superiores de precisão. Tanto as correções analíticas quanto as baseadas em bootstrap podem ser altamente eficazes.

A aplicação do método bootstrap na correção de viés tem sido explorada em diferentes contextos estatísticos, revelando seu potencial de melhoria em estimadores de máxima verossimilhança. O estudo de [Cribari-Neto e Vasconcellos \(2002\)](#) usa o método para corrigir o viés em modelos de regressão baseados na distribuição beta, adequada para modelar variáveis contínuas restritas ao intervalo $(0,1)$. A abordagem foi comparada com a clássica correção analítica de segunda ordem proposta por Cox e Snell, avaliando o desempenho de versões corretivas e preventivas do viés. Os resultados indicaram que, embora o estimador obtido por bootstrap apresentasse maior erro quadrático médio devido à sua variabilidade, ele foi eficaz na redução do viés.

No estudo conduzido por [Lemonte, Simas e Cribari-Neto \(2008\)](#) foi avaliada a eficiência do bootstrap em sua versão paramétrica e não paramétrica na correção de viés de estimadores aplicados à distribuição Birnbaum-Saunders, comumente utilizada para modelar tempos de falha. O estudo demonstrou que os esquemas de correção baseados em bootstrap são eficazes mesmo nos menores tamanhos amostrais considerados no estudo. Foram analisados três estimadores de Bootstrap, correção de viés constante (CBC), linear (LBC) e o baseado diretamente na reamostragem dos estimadores bootstrap. Os resultados numéricos indicaram superioridade do CBC paramétrico, sendo este recomendado pelos autores para o caso da distribuição Birnbaum-Saunders (mais detalhes sobre o método CBC e LBC podem ser encontrados em [MacKinnon e Jr \(1998\)](#)).

A robustez do bootstrap tem sido evidenciada em outras aplicações de distintas distribuições ao longos dos anos. Por exemplo, [Lagos-Álvarez, Jiménez-Gamero e Alba-Fernández \(2011\)](#) mostraram que o bootstrap paramétrico e o jackknife superam correções analíticas e modificações nas equações de verossimilhança em termos de eficácia para pequenas amostras no caso da distribuição logística generalizada tipo I. Conclusão similar foi apresentada para a distribuição Lomax ([GILES; FENG; GODWIN, 2013](#)). No trabalho de [Dore et al. \(2016\)](#) foram desenvolvidas correções de viés analíticas e por bootstrap para o EMV dos autovalores da matriz de parâmetros da distribuição de Bingham complexa, e ambas demonstraram relevante redução do erro quadrático médio dos EMVs. O bootstrap paramétrico demonstra bons resultados, sobretudo em pequenas amostras e em situações de baixa confiabilidade e não normalidade, porém, embora o bootstrap seja amplamente efetivo na correção de viés, sua performance está intrinsecamente ligada ao equilíbrio entre viés e variância ([MAGALHÃES; GALLARDO; BOURGUIGNON, 2021](#); [DHAENE; ROSSEEL, 2022](#); [MEDEIROS; ARAÚJO; BOURGUIGNON, 2023](#)). Estes estudos apontam que, enquanto reduz o viés, a variabilidade introduzida pelo método pode aumentar o erro quadrático médio. Ainda assim, a simplicidade de implementação, aplicabilidade

em modelos complexos e consistência assintótica tornam o bootstrap uma relevante ferramenta para aprimorar estimadores em estatística aplicada.

3.2.3.1 Estimadores Corrigidos

Considerando inicialmente o esquema paramétrico do método bootstrap, baseado na amostra aleatória $Y_i = (Y_1, \dots, Y_q)$ com função densidade definida na Equação (2.1), $\beta = \text{vec}(\beta_1, \dots, \beta_q)$ o vetor de parâmetros do modelo de regressão Dirichlet e a estimativa do EMV dada por $\hat{\beta}$. A partir da amostra original de dados gera-se um número B de amostras independentes da distribuição Dirichlet, $Y_1^*, \dots, Y_B^*$, com a função densidade $f(x; \hat{\beta})$. Para cada amostra paramétrica de bootstrap (Y_b^*), em que $b = 1, \dots, B$, calcula-se um novo estimador $\hat{\beta}_b^*$. A estimativa do viés para o EMV $\hat{\beta}_b^*$ é dado por

$$\hat{B}(\hat{\beta}_b^*) = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b^* - \beta \quad (3.23)$$

onde $\hat{\beta}_b^*$ é o EMV de β obtido na b -ésima amostra de bootstrap. Com isso, o estimador corrigido por bootstrap paramétrico $\hat{\beta}_{boot}$ é dado por $\hat{\beta}_{boot} = \hat{\beta} - \hat{B}(\hat{\beta}_b^*)$, isto é

$$\hat{\beta}_{boot} = 2\hat{\beta} - \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b^* \quad (3.24)$$

O procedimento não-paramétrico é análogo a versão paramétrica, mas a reamostragem é feita a partir da distribuição empírica da amostra original, dita $\hat{F}$, sem assumir qualquer suposição paramétrica para as amostras de bootstrap e amostrando os dados com reposição (CORDEIRO; CRIBARI-NETO et al., 2014). A função da distribuição empírica da amostra original é dada por

$$\hat{F} = \frac{\#\{y_i \leq y\}}{n}$$

onde se atribui a cada valor da amostra (y_i) igual probabilidade n^{-1} , e $\#\{A\}$ denota o número de elementos em A (DAVISON; HINKLEY, 1997, pg. 12).

O método não paramétrico para correção do viés pode ser descrito como:

- Considere $Y = (Y_1, \dots, Y_q)$ um vetor aleatório com função densidade dada em (2.1), estima-se o EMV para β , o estimador $\hat{\beta}$ dado na Equação (2.4);
- A partir da amostra original, seleciona-se B novas amostras com reposição $Y_1^*, \dots, Y_B^*$;
- Para cada amostra de bootstrap $b = 1, \dots, B$ calcula-se o EMV $\hat{\beta}_b^*$;
- Com a coleção das estimativas $\{\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*, \dots, \hat{\beta}_B^*\}$ calcula-se o estimador esperado de bootstrap;

- A correção do estimador é dada por $\hat{\beta}_{boot}$ como calculado anteriormente na Equação (3.24).

O procedimento descrito realiza a correção do estimador pela redução do viés de segunda ordem. A vantagem é que o método não exige derivações analíticas do viés de segunda ordem ou da matriz de informação de Fisher, como nos métodos analíticos (VASCONCELLOS; DOURADO, 2008)

3.3 Avaliações Numéricas

Neste capítulo, é realizada a avaliação numérica da correção do viés por meio da análise experimental, sendo considerado o modelo de regressão de Dirichlet com parametrização dada na Equação 2.3. Com a finalidade de avaliar o desempenho dos estimadores de máxima verossimilhança e sua versão corrigida, foi verificado o viés médio, calculado para um parâmetro qualquer θ , como

$$\frac{\sum_{i=1}^R (\hat{\theta}_i - \theta_i)}{R}$$

a fim de identificar a magnitude da distorção sistemática e o erro quadrático médio padronizado (EQM) dos parâmetros, como medida de precisão dos estimadores, calculado genericamente pela Equação (3.25).

$$EQM(\hat{\theta}) = \sqrt{\frac{1}{R} \sum_{i=1}^R (\hat{\theta}_i - \theta)^2} \quad (3.25)$$

onde R representa o número de réplicas de Monte Carlo e $\hat{\theta}_i$ é a i -ésima estimativa no procedimento de Monte Carlo.

O experimento de simulação foi conduzido por meio de um algoritmo de Monte Carlo, utilizado para gerar sucessivas amostras aleatórias dos dados sob diferentes configurações de modelos e tamanhos amostrais.

Em cada cenário do esquema de bootstrap, foram consideradas 5.000 réplicas de Monte Carlo, às quais se aplicou o procedimento de correção baseado em 200 amostras de bootstrap, em consonância com a literatura revisada (vide seção 3.1), tanto na versão paramétrica quanto na versão não paramétrica. Assim, cada combinação de modelo e tamanho amostral envolveu um total de 1.000.000 de iterações. Segundo (DAVISON; HINKLEY, 1997) para muitos propósitos práticos, um número entre 100 e 200 réplicas de bootstrap pode ser suficiente. Para o esquema do método de Cox-Snell foi considerado 10.000 réplicas de Monte Carlo.

A modelagem assumiu que a variável resposta é uma quantidade composicional formada por três componentes,

$$Y = (Y_1, Y_2, Y_3)$$

as quais foram relacionadas às variáveis explicativas contínuas x_1, x_2 e x_3 , portanto $p = q = 3$. Para cada componente, os parâmetros foram definidos a partir de uma estrutura linear base, representada pelo modelo definido anteriormente na Equação (3.22), com a adição do intercepto, como segue:

$$Y_i = \begin{cases} Y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ Y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ Y_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \end{cases} \quad (3.26)$$

Os valores das variáveis explicativas foram gerados aleatoriamente segundo uma distribuição uniforme padrão $X \sim U(0, 1)$

Para analisar o impacto do tamanho amostral sobre o viés e a precisão das estimativas, foram considerados os seguintes tamanhos de amostra:

$$n \in \{20, 30, 40, 50, 60, 100, 500, 1000\}$$

Além dos modelos com especificação linear como anteriormente definidos, o estudo incorporou uma segunda classe de modelos: os modelos polinomiais, cujo objetivo foi avaliar o comportamento dos estimadores sob estruturas mais complexas de relação entre as variáveis explicativas e a resposta composicional.

A construção das variáveis explicativas polinomiais baseou-se na geração inicial de um preditor contínuo x , definido como uma sequência no intervalo $(0,1)$. Especificamente, para uma amostra de tamanho n , o preditor base foi gerado como:

$$x = \frac{1}{n+1}, \frac{2}{n+1}, \dots, \frac{n}{n+1}$$

Esse preditor serviu como base para a construção de polinômios de ordem superior.

Para a modelagem polinomial também foram considerados três preditores (x_1, x_2, x_3) , definidos a partir das potências sucessivas de x , da forma x^k , $k \in (1, \dots, p)$, tal que p é o número de variáveis preditoras.

Assim, o vetor de covariáveis utilizado nos modelos polinomiais assumiu a forma:

$$X = (1, x, x^2, x^3),$$

onde o primeiro elemento corresponde ao termo constante do modelo.

As simulações foram concluídas usando a versão 4.4.2 do software estatístico R (2024) para sistema operacional Windows (TEAM, 2024). A função de log-verossimilhança foi maximizada utilizando o método de otimização não-linear quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) para a estimação dos parâmetros da Regressão Dirichlet, conforme implementado em Maier (2014), usando o pacote *DirichletReg*.

As simulações de Monte Carlo realizadas nessa etapa do estudo foram realizadas em um computador com sistema operacional Windows 11 (64 bits), equipado com processador Intel Core i7-13650HX (13ª geração, 20 processadores lógicos), 16 GiB de memória RAM, unidade de armazenamento SSD de 512 GB e placa de vídeo NVIDIA GeForce RTX.

3.3.1 Resultados do Experimento

Avalia-se nessa seção o comportamento do viés do EMV de β , seu erro quadrático médio (EQM), e a correspondente estimativa de viés após a aplicação do procedimento de correção. A notação com subscrito (c) é utilizada para identificar o método empregado na correção do viés.

Inicialmente, analisam-se os modelos linearmente especificados (vide Eq. (3.3)). A Tabela 2 apresenta os resultados obtidos com a aplicação do bootstrap paramétrico. Observa-se a presença evidente de viés no estimador de máxima verossimilhança (EMV) no contexto da regressão Dirichlet, especialmente para os tamanhos amostrais reduzidos ($n \leq 30$). Por exemplo, para $n = 20$, o viés médio do estimador é aproximadamente seis vezes maior que o observado para $n = 30$ e onze vezes maior que o estimador para $n = 60$, evidenciando o impacto do tamanho amostral na estimação dos parâmetros. Verifica-se ainda que o viés tende a zero à medida que o tamanho amostral aumenta, apresentando média de aproximadamente 0.003 e desvio-padrão de ± 0.0019 para $n = 1000$.

A aplicação do método de reamostragem indica que a correção por bootstrap paramétrico reduz o viés dos estimadores na regressão Dirichlet sem comprometer a precisão. Por exemplo, considerando o experimento com $n = 20$, o viés médio, em módulo, dos estimadores corrigidos pelo bootstrap paramétrico foi de aproximadamente 0,19 ($\pm 0,03$ de desvio-padrão), enquanto o viés médio do EMV foi de 0,58 ($\pm 0,09$). Nesse cenário, a técnica resultou em uma redução média do viés de aproximadamente 67,7%. O erro quadrático médio (EQM) para o estimador com viés corrigido foi, em média, 1,17, enquanto para o EMV foi de 1,68, indicando que a variabilidade das estimativas também foi reduzida após a aplicação da correção.

Com o aumento do tamanho amostral ($n = 500$ e $n = 1000$), tanto o viés quanto o EQM diminuem naturalmente, sendo que a correção ainda proporcionou ganhos marginais adicionais. Nesse cenário, alguns estimadores apresentaram viés praticamente nulo.

Entretanto, duas situações não registraram redução no viés médio: os estimadores associados aos parâmetros β_{01} e β_{02} quando $n = 60$. Nesses casos, observou-se um leve

Tabela 2 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Bootstrap Paramétrico (modelo linearmente especificado).

Parâmetros	N = 20				N = 30				N = 40				N = 50			
	Viés	Viés _{bp}	EQM	EQM _{bp}	Viés	Viés _{bp}	EQM	EQM _{bp}	Viés	Viés _{bp}	EQM	EQM _{bp}	Viés	Viés _{bp}	EQM	EQM _{bp}
β_{01}	-0.452	0.219	1.707	1.177	0.072	0.017	0.657	0.603	0.073	-0.003	0.504	0.474	0.037	0.006	0.448	0.427
β_{02}	-0.453	0.218	1.712	1.182	0.070	0.014	0.654	0.601	0.072	-0.004	0.503	0.473	0.032	0.001	0.451	0.431
β_{03}	-0.460	0.211	1.708	1.174	0.071	0.015	0.651	0.597	0.070	-0.006	0.501	0.472	0.032	0.001	0.449	0.428
β_{11}	0.541	-0.143	1.528	1.090	0.142	-0.033	0.983	0.854	0.173	-0.009	0.668	0.591	0.111	-0.018	0.561	0.521
β_{12}	0.543	-0.140	1.535	1.097	0.143	-0.032	0.983	0.854	0.179	-0.003	0.675	0.596	0.112	-0.016	0.562	0.522
β_{13}	0.545	-0.138	1.525	1.084	0.147	-0.028	0.976	0.847	0.175	-0.007	0.671	0.593	0.115	-0.013	0.561	0.520
β_{21}	0.700	-0.206	1.731	1.157	0.164	-0.016	0.769	0.699	0.056	-0.006	0.679	0.626	-0.016	-0.005	0.618	0.581
β_{22}	0.707	-0.199	1.735	1.158	0.168	-0.012	0.775	0.704	0.055	-0.008	0.677	0.625	-0.014	-0.004	0.615	0.578
β_{23}	0.707	-0.199	1.728	1.150	0.162	-0.018	0.771	0.702	0.055	-0.007	0.674	0.621	-0.013	-0.002	0.619	0.582
β_{31}	0.603	-0.185	1.748	1.248	0.012	-0.007	0.964	0.849	-0.026	0.018	0.761	0.691	0.102	0.006	0.737	0.681
β_{32}	0.597	-0.192	1.740	1.243	0.013	-0.007	0.958	0.842	-0.027	0.017	0.768	0.698	0.107	0.011	0.739	0.682
β_{33}	0.606	-0.182	1.741	1.238	0.013	-0.007	0.953	0.838	-0.023	0.021	0.767	0.696	0.103	0.006	0.738	0.682

Parâmetros	N = 60				N = 100				N = 500				N = 1000			
	Viés	Viés _{bp}	EQM	EQM _{bp}	Viés	Viés _{bp}	EQM	EQM _{bp}	Viés	Viés _{bp}	EQM	EQM _{bp}	Viés	Viés _{bp}	EQM	EQM _{bp}
β_{01}	0.011	-0.012	0.466	0.447	0.008	0.002	0.315	0.309	0.002	-0.001	0.143	0.143	0.001	-0.001	0.105	0.105
β_{02}	0.011	-0.012	0.469	0.451	0.006	-0.001	0.317	0.311	0.002	-0.001	0.143	0.143	0.002	0.000	0.105	0.105
β_{03}	0.013	-0.010	0.466	0.447	0.007	0.000	0.315	0.309	0.004	0.000	0.142	0.141	0.001	-0.001	0.104	0.104
β_{11}	0.070	0.000	0.468	0.445	0.037	0.003	0.367	0.356	0.010	0.003	0.159	0.158	0.006	0.003	0.112	0.112
β_{12}	0.068	-0.001	0.468	0.446	0.038	0.004	0.366	0.355	0.010	0.003	0.160	0.159	0.006	0.003	0.113	0.113
β_{13}	0.064	-0.006	0.470	0.449	0.037	0.004	0.370	0.359	0.008	0.002	0.157	0.157	0.006	0.003	0.113	0.112
β_{21}	0.038	-0.002	0.532	0.509	0.017	-0.010	0.402	0.391	0.005	-0.001	0.158	0.158	0.002	0.000	0.113	0.112
β_{22}	0.042	0.002	0.531	0.508	0.020	-0.007	0.402	0.392	0.005	-0.001	0.158	0.158	0.002	-0.001	0.113	0.113
β_{23}	0.043	0.004	0.532	0.508	0.020	-0.007	0.401	0.390	0.005	-0.002	0.159	0.158	0.002	0.000	0.113	0.113
β_{31}	0.081	0.022	0.603	0.569	0.058	0.002	0.385	0.372	0.004	0.000	0.157	0.157	0.002	0.000	0.114	0.114
β_{32}	0.082	0.022	0.606	0.572	0.057	0.001	0.390	0.376	0.004	0.000	0.156	0.156	0.002	-0.001	0.115	0.115
β_{33}	0.080	0.020	0.606	0.572	0.056	0.000	0.386	0.373	0.003	0.000	0.157	0.157	0.002	0.000	0.115	0.115

Fonte: Produzido pelos autores.

acréscimo no viés relativo, de aproximadamente 9,09%. Ainda assim, verificou-se redução do erro quadrático médio nesse cenário, bem como para os demais parâmetros, em todos os tamanhos amostrais.

Em seguida, avalia-se a efetividade do método de bootstrap não paramétrico também para o caso em que o modelo é linearmente especificado. Os resultados da simulação estão apresentados na Tabela 3. A técnica não apresentou convergência numérica quando $n = 20$, o que pode ser atribuído à instabilidade do processo de reamostragem em amostras muito pequenas. Dessa forma, os resultados foram avaliados apenas para tamanhos amostrais $n \geq 30$.

A primeira conclusão observada é que não houve redução do viés relativo para o menor tamanho amostral considerado ($n = 30$). Ainda assim, a variabilidade das estimativas mostrou-se inferior à observada para o EMV.

De modo semelhante ao observado anteriormente no caso paramétrico, o viés médio do EMV tende a zero à medida que o tamanho amostral aumenta, atingindo magnitude ainda menor quando $n = 1000$. Como os valores obtidos foram extremamente pequenos nesse cenário, adotou-se a notação científica para facilitar a apresentação dos resultados. O viés relativo médio quando $n = 1000$ foi de aproximadamente 7.944×10^{-11} , com desvio-padrão de $\pm 5,62 \times 10^{-11}$.

Considerando o experimento com $n = 40$, o viés médio, em módulo, dos estimadores corrigidos pelo bootstrap não paramétrico foi de aproximadamente 0,009 ($\pm 0,04$), enquanto o viés médio do EMV foi de 0,082 ($\pm 0,02$). Nesse cenário, a técnica resultou em uma redução média do viés de aproximadamente 61,21%. Esse resultado indica um desempenho inferior ao observado para o bootstrap paramétrico, visto que, para o mesmo tamanho amostral, este apresentou redução média do viés de aproximadamente 88,81%.

Para $n = 40$, o erro quadrático médio (EQM) do estimador com viés corrigido foi, em média, 0,66, enquanto para o EMV foi de 0,62. De forma semelhante ao observado para o viés médio, a magnitude da redução do EQM obtida pelo bootstrap não paramétrico foi inferior à observada na versão paramétrica.

Alguns outros casos não registraram redução no viés médio. Isso ocorreu para os estimadores associados aos parâmetros β_{11} , β_{12} , β_{13} e β_{23} quando $n = 1000$, bem como para β_{31} e β_{32} quando $n = 500$. Além disso, considerando que o viés já é naturalmente reduzido em tamanhos amostrais maiores, a contribuição adicional da técnica tornou-se limitada nesses cenários.

A efetividade do método de reamostragem é comparada ao método analítico. Para tanto, avaliam-se os resultados da simulação obtidos pelo método de Cox–Snell para um modelo com intercepto, com o objetivo de permitir a comparação com os resultados do bootstrap previamente discutidos. Tais resultados estão apresentados na Tabela 4.

No caso analítico, o viés médio dos estimadores apresentou magnitude semelhante àquela observada para os estimadores obtidos pelas técnicas de bootstrap. Esse resultado indica que

Tabela 4 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de CoxSnell.

Parâmetros	$N = 20$					$N = 30$					$N = 40$					$N = 50$				
	B		B_{cs}		EQM	B		B_{cs}		EQM	B		B_{cs}		EQM	B		B_{cs}		EQM
	B	B_{cs}	B	B_{cs}	EQM	B	B_{cs}	B	B_{cs}	EQM	B	B_{cs}	B	B_{cs}	EQM	B	B_{cs}	B	B_{cs}	EQM
β_{01}	-0.005	-0.124	1.577	1.411	0.027	-0.047	0.859	0.795	0.040	-0.015	0.664	0.626	0.027	-0.016	0.547	0.027	-0.016	0.547	-0.016	0.547
β_{02}	-0.005	-0.125	1.578	1.413	0.028	-0.046	0.861	0.797	0.043	-0.012	0.665	0.627	0.030	-0.013	0.547	0.030	-0.013	0.547	-0.013	0.547
β_{03}	-0.004	-0.124	1.580	1.415	0.032	-0.043	0.858	0.794	0.042	-0.013	0.667	0.629	0.028	-0.015	0.547	0.028	-0.015	0.547	-0.015	0.547
β_{11}	0.300	0.169	1.867	1.642	0.151	0.064	0.987	0.896	0.080	0.016	0.745	0.695	0.064	0.012	0.616	0.064	0.012	0.616	0.012	0.616
β_{12}	0.301	0.170	1.863	1.638	0.153	0.066	0.993	0.902	0.079	0.015	0.745	0.694	0.065	0.013	0.616	0.065	0.013	0.616	0.013	0.616
β_{13}	0.297	0.166	1.869	1.644	0.152	0.065	0.988	0.897	0.080	0.016	0.744	0.694	0.063	0.011	0.614	0.063	0.011	0.614	0.011	0.614
β_{21}	0.293	0.162	1.845	1.622	0.148	0.061	0.988	0.898	0.097	0.033	0.742	0.692	0.070	0.018	0.618	0.070	0.018	0.618	0.018	0.618
β_{22}	0.290	0.159	1.847	1.624	0.147	0.060	0.990	0.899	0.093	0.029	0.739	0.689	0.065	0.013	0.618	0.065	0.013	0.618	0.013	0.618
β_{23}	0.294	0.163	1.842	1.620	0.145	0.058	0.986	0.896	0.093	0.028	0.745	0.695	0.069	0.017	0.619	0.069	0.017	0.619	0.017	0.619
β_{31}	0.277	0.151	1.867	1.644	0.127	0.040	0.986	0.898	0.080	0.015	0.753	0.702	0.071	0.019	0.612	0.071	0.019	0.612	0.019	0.612
β_{32}	0.279	0.152	1.862	1.638	0.124	0.038	0.988	0.900	0.079	0.015	0.756	0.704	0.069	0.017	0.614	0.069	0.017	0.614	0.017	0.614
β_{33}	0.277	0.151	1.871	1.648	0.122	0.035	0.990	0.901	0.079	0.014	0.757	0.706	0.071	0.019	0.615	0.071	0.019	0.615	0.019	0.615

Parâmetros	$N = 60$					$N = 100$					$N = 500$					$N = 1000$				
	B		B_{cs}		EQM	B		B_{cs}		EQM	B		B_{cs}		EQM	B		B_{cs}		EQM
	B	B_{cs}	B	B_{cs}	EQM	B	B_{cs}	B	B_{cs}	EQM	B	B_{cs}	B	B_{cs}	EQM	B	B_{cs}	B	B_{cs}	EQM
β_{01}	0.028	-0.007	0.490	0.471	0.018	-0.003	0.346	0.338	0.006	0.002	0.144	0.143	0.000	-0.002	0.101	0.000	-0.002	0.101	-0.002	0.101
β_{02}	0.024	-0.011	0.487	0.468	0.014	-0.006	0.344	0.337	0.006	0.002	0.143	0.142	0.000	-0.002	0.101	0.000	-0.002	0.101	-0.002	0.101
β_{03}	0.025	-0.010	0.487	0.468	0.016	-0.005	0.348	0.340	0.006	0.002	0.144	0.143	0.000	-0.002	0.101	0.000	-0.002	0.101	-0.002	0.101
β_{11}	0.054	0.010	0.546	0.521	0.036	0.009	0.385	0.375	0.004	-0.001	0.159	0.158	0.003	0.001	0.112	0.003	0.001	0.112	0.001	0.111
β_{12}	0.056	0.012	0.545	0.520	0.034	0.008	0.384	0.373	0.004	-0.002	0.158	0.157	0.003	0.001	0.111	0.003	0.001	0.111	0.001	0.111
β_{13}	0.053	0.009	0.544	0.519	0.034	0.008	0.386	0.375	0.004	-0.001	0.158	0.157	0.004	0.001	0.111	0.004	0.001	0.111	0.001	0.111
β_{21}	0.057	0.014	0.546	0.521	0.026	0.000	0.387	0.376	0.004	-0.001	0.159	0.158	0.004	0.002	0.110	0.004	0.002	0.110	0.002	0.110
β_{22}	0.057	0.013	0.542	0.516	0.028	0.002	0.386	0.375	0.004	-0.001	0.159	0.158	0.004	0.002	0.111	0.004	0.002	0.111	0.002	0.111
β_{23}	0.059	0.015	0.542	0.517	0.029	0.003	0.388	0.377	0.004	-0.001	0.159	0.158	0.005	0.002	0.111	0.005	0.002	0.111	0.002	0.111
β_{31}	0.048	0.005	0.547	0.522	0.028	0.002	0.384	0.374	0.004	-0.001	0.158	0.157	0.004	0.001	0.110	0.004	0.001	0.110	0.001	0.110
β_{32}	0.053	0.010	0.545	0.519	0.033	0.007	0.384	0.373	0.004	-0.001	0.158	0.158	0.004	0.001	0.110	0.004	0.001	0.110	0.001	0.110
β_{33}	0.054	0.011	0.545	0.520	0.030	0.004	0.383	0.372	0.004	-0.001	0.158	0.157	0.004	0.001	0.111	0.004	0.001	0.111	0.001	0.111

Fonte: Produzido pelos autores.

os métodos de reamostragem não introduzem viés adicional em relação à abordagem analítica. Como exceção, observa-se que, quando $n = 20$, o viés médio obtido pela correção analítica apresentou magnitude inferior à observada nos métodos de bootstrap. Também evidencia-se que para o intercepto do modelo não foi observada redução de viés quando $n \leq 30$. Para os demais parâmetros a redução do viés foi de aproximadamente 68% ($\pm 10,12\%$)

Em média, a correção analítica reduziu o viés relativo dos estimadores em aproximadamente 30,8% quando $n = 20$. Nesse cenário, observa-se que o bootstrap paramétrico apresentou desempenho superior na redução do viés. Entretanto, à medida que o tamanho amostral aumenta, os valores de viés relativo obtidos pelas diferentes técnicas tornam-se próximos, evidenciando comportamento semelhante entre os métodos. Esse comportamento é esperado uma vez que em tamanhos amostrais maiores o viés é naturalmente reduzido, corroborando com a literatura sobre a correção de viés (ver seção 3.1).

Verifica-se também que o erro quadrático médio (EQM) dos estimadores corrigidos pelo método de Cox–Snell tende a ser inferior ao observado para o EMV, indicando melhoria na precisão das estimativas, resultado esperado conforme a literatura.

Assim como observado nos métodos de bootstrap, para tamanhos amostrais grandes ($n = 500$ e $n = 1000$) tanto o viés quanto o EQM tornam-se naturalmente muito pequenos, refletindo as propriedades assintóticas do estimador de máxima verossimilhança. Esse resultado confirma que a correção de Cox–Snell não introduz instabilidade adicional nas estimativas e que, em amostras grandes, as diferenças entre os estimadores corrigidos e o EMV tornam-se praticamente desprezíveis.

Registra-se ainda que nem todos os estimadores apresentaram redução de viés após a aplicação da correção analítica. Esse comportamento foi observado para os estimadores associados ao intercepto do modelo quando $n = 20$ e $n = 1000$, isto é, nos extremos do tamanho amostral considerado. Ainda assim, em todos os casos foi registrado melhoria na precisão. O EQM dos estimadores foi inferior após a correção.

Na sequência procede-se à avaliação da segunda classe de modelos, composta por especificações com termos polinomiais. Os resultados da simulação Monte Carlo para esse cenário estão descritos nas Tabelas 5 e 6. Nesse experimento, avalia-se a efetividade dos métodos de reamostragem diante do aumento da complexidade do modelo, representado pela inclusão de termos polinomiais de maior grau.

Os resultados evidenciam um aumento no viés médio do estimador de máxima verossimilhança (EMV) com a introdução dos termos polinomiais no modelo. No caso do bootstrap paramétrico, o viés dos estimadores associados aos parâmetros das covariáveis mostrou-se significativamente superior ao observado para o intercepto. Entretanto, observa-se que o aumento da ordem do termo polinomial não implica necessariamente aumento proporcional na magnitude do viés médio. Por exemplo, no experimento com $n = 20$, o viés médio do estimador do parâmetro

β_{11} foi de aproximadamente $-20,96$, enquanto o viés do parâmetro de ordem superior β_{21} foi de $-11,65$. De forma semelhante, para o experimento com $n = 30$, o viés do estimador associado a β_{12} foi de $-13,32$, enquanto para β_{22} observou-se magnitude inferior, com valor aproximado de $3,82$.

A técnica de bootstrap paramétrico mostrou-se efetiva também nesse contexto. Para o menor tamanho amostral considerado ($n = 20$), a redução média do viés foi de aproximadamente $40,05\%$, sendo ainda mais expressiva para os demais tamanhos amostrais. Para o tamanho amostral imediatamente superior ($n = 30$), por exemplo, a redução média do viés atingiu aproximadamente $79,53\%$.

O método também foi efetivo na redução do erro quadrático médio (EQM) dos estimadores. Contudo, o EQM associado ao EMV apresentou valores elevados em todos os tamanhos amostrais analisados. A diferença entre os valores corrigidos e não corrigidos diminui progressivamente com o aumento do tamanho amostral. O maior ganho em termos de precisão foi observado para $n = 20$, enquanto, para amostras maiores, os valores de EQM entre os estimadores tornam-se progressivamente mais próximos.

O comportamento da aplicação da técnica de bootstrap não paramétrico no modelo com termos polinomiais mostra-se semelhante ao observado no modelo linearmente especificado. Para o menor tamanho amostral considerado ($n = 30$), não se verifica redução efetiva do viés médio. Nesse caso, o viés médio do EMV foi de aproximadamente $8,86 (\pm 4,19)$ e, após a aplicação da correção por bootstrap não paramétrico, passou a registrar média de $17,24 (\pm 13,20)$. Entretanto, para os demais tamanhos amostrais observa-se redução do viés médio dos estimadores.

Quanto ao erro quadrático médio (EQM), embora se observe redução na maioria dos casos, para amostras maiores ($n \geq 30$) os valores obtidos para os estimadores do EMV e para os estimadores corrigidos tornam-se progressivamente mais próximos. Esse comportamento reflete o fato de que, à medida que o tamanho amostral aumenta, o viés do estimador de máxima verossimilhança tende naturalmente a diminuir, reduzindo também a contribuição adicional proporcionada pela correção de viés baseada em reamostragem.

Tabela 5 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Bootstrap Paramétrico (especificação polinomial).

	N = 20					N = 30					N = 40					N = 50				
Parâmetros	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}
β_{01}	8.640	-0.835	7.311	4.903	4.793	-0.244	4.142	3.466	3.389	-0.063	3.077	2.708	2.528	-0.082	2.520	2.288				
β_{02}	8.768	-0.700	7.432	5.016	4.754	-0.282	4.094	3.425	3.362	-0.087	3.079	2.712	2.612	0.000	2.574	2.336				
β_{03}	8.510	-0.959	7.351	4.977	4.881	-0.173	4.125	3.435	3.431	-0.020	3.091	2.713	2.624	0.012	2.545	2.304				
β_{11}	-20.965	-2.103	56.781	39.655	-13.689	0.293	31.974	28.220	-10.354	-0.228	24.347	22.540	-7.266	0.425	20.383	19.369				
β_{12}	-22.029	-3.224	57.436	40.338	-13.321	0.630	31.640	27.916	-10.331	-0.209	24.556	22.761	-7.963	-0.310	20.652	19.617				
β_{13}	-20.457	-1.662	57.124	40.176	-14.209	-0.145	31.884	28.089	-10.790	-0.668	24.494	22.640	-7.965	-0.244	20.199	19.177				
β_{21}	-11.653	21.418	131.535	90.860	4.526	2.684	72.462	64.066	6.142	1.807	55.380	51.546	3.785	-0.123	46.561	44.455				
β_{22}	-9.513	23.673	132.340	91.974	3.820	2.037	71.716	63.322	5.895	1.549	56.003	52.190	5.053	1.234	46.896	44.791				
β_{23}	-12.427	20.837	132.104	91.831	5.804	3.840	72.171	63.742	6.697	2.396	55.806	51.895	5.121	1.141	45.960	43.921				
β_{31}	39.933	-22.793	90.523	61.007	11.473	-3.474	48.336	42.325	5.395	-1.877	36.647	33.991	4.435	-0.363	30.702	29.241				
β_{32}	38.628	-24.165	90.774	61.562	11.859	-3.114	47.840	41.800	5.668	-1.589	37.071	34.416	3.810	-1.035	30.820	29.379				
β_{33}	40.280	-22.588	90.780	61.530	10.578	-4.313	48.033	42.036	5.270	-2.038	36.956	34.251	3.690	-1.061	30.300	28.911				
	N = 60					N = 100					N = 500					N = 1000				
Parâmetros	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}	B	B _{bp}	EQM	EQM _{bp}
β_{01}	1.936	-0.165	2.142	1.986	1.169	-0.019	1.547	1.476	0.184	-0.038	0.618	0.613	0.112	0.001	0.424	0.424				
β_{02}	1.876	-0.220	2.183	2.033	1.296	0.104	1.563	1.482	0.209	-0.013	0.612	0.607	0.095	-0.016	0.424	0.423				
β_{03}	1.927	-0.175	2.187	2.032	1.144	-0.041	1.508	1.434	0.211	-0.013	0.614	0.610	0.125	0.012	0.423	0.422				
β_{11}	-4.888	1.304	17.469	16.821	-3.657	-0.094	12.737	12.468	-0.249	0.423	5.230	5.214	-0.333	0.008	3.616	3.622				
β_{12}	-4.149	2.011	17.954	17.309	-4.476	-0.928	12.855	12.557	-0.515	0.157	5.207	5.196	-0.273	0.067	3.640	3.648				
β_{13}	-4.581	1.626	17.770	17.114	-3.665	-0.129	12.513	12.219	-0.514	0.166	5.192	5.180	-0.520	-0.174	3.602	3.605				
β_{21}	0.951	-2.470	40.309	38.929	3.006	0.629	29.162	28.642	-0.414	-0.925	12.034	12.007	0.235	-0.033	8.306	8.328				
β_{22}	-0.422	-3.789	41.340	39.921	4.550	2.234	29.385	28.831	0.089	-0.417	12.009	11.998	0.230	-0.034	8.393	8.417				
β_{23}	0.165	-3.297	40.713	39.296	3.282	0.967	28.778	28.214	0.053	-0.471	11.930	11.911	0.710	0.440	8.281	8.295				
β_{31}	4.852	1.321	26.735	25.744	0.974	-0.586	19.135	18.792	0.814	0.590	7.879	7.861	0.122	0.018	5.419	5.434				
β_{32}	5.547	1.985	27.312	26.286	0.105	-1.499	19.268	18.916	0.545	0.318	7.877	7.870	0.069	-0.037	5.478	5.495				
β_{33}	5.407	1.903	26.886	25.861	0.741	-0.858	18.941	18.580	0.587	0.370	7.817	7.805	-0.192	-0.298	5.410	5.422				

Fonte: Produzido pelos autores.

Tabela 6 – Viés médio, Erro quadrático médio (EQM) e versões corrigidas pelo método de Bootstrap Não Paramétrico (especificação polinomial).

Parâmetros	N = 30						N = 40						N = 50						N = 60					
	B		B _{bnp}		EQM		B		B _{bnp}		EQM		B		B _{bnp}		EQM		B		B _{bnp}		EQM	
	B	B _{bnp}	B	B _{bnp}	EQM	EQM _{bnp}	B	B _{bnp}	B	B _{bnp}	EQM	EQM _{bnp}	B	B _{bnp}	B	B _{bnp}	EQM	EQM _{bnp}	B	B _{bnp}	B	B _{bnp}	EQM	EQM _{bnp}
β_{01}	4.867	-7.315	4.114	4.676	4.676	2.879	3.587	-0.896	3.087	2.879	2.879	2.879	2.659	-0.099	2.554	2.364	2.106	0.078	2.232	2.106	0.078	2.232	2.094	2.094
β_{02}	4.732	-7.456	4.098	4.715	4.715	2.857	3.444	-1.048	3.055	2.857	2.857	2.857	2.621	-0.139	2.540	2.352	2.078	0.043	2.197	2.078	0.043	2.197	2.058	2.058
β_{03}	4.688	-7.485	4.072	4.672	4.672	2.922	3.526	-0.937	3.121	2.922	2.922	2.922	2.545	-0.201	2.530	2.357	1.927	-0.112	2.192	1.927	-0.112	2.192	2.072	2.072
β_{11}	-15.402	26.123	32.405	33.889	33.889	23.695	-11.216	1.680	24.200	23.695	23.695	23.695	-8.174	-0.512	20.470	19.882	-6.850	-1.346	17.844	-6.850	-1.346	17.844	17.421	17.421
β_{12}	-13.864	27.760	32.096	34.064	34.064	23.672	-10.206	2.706	24.074	23.672	23.672	23.672	-7.967	-0.232	20.316	19.709	-6.755	-1.264	17.736	-6.755	-1.264	17.736	17.278	17.278
β_{13}	-13.149	28.179	31.939	33.810	33.810	23.828	-10.753	2.012	24.314	23.828	23.828	23.828	-7.656	-0.097	20.361	19.779	-5.289	0.280	17.714	-5.289	0.280	17.714	17.373	17.373
β_{21}	8.465	1.010	73.942	74.677	74.677	54.284	7.151	6.704	55.084	54.284	54.284	54.284	5.196	3.455	46.699	45.714	5.042	3.261	40.446	5.042	3.261	40.446	39.763	39.763
β_{22}	5.249	-2.566	73.272	74.776	74.776	54.336	5.677	5.222	54.944	54.336	54.336	54.336	4.618	2.693	46.384	45.326	5.099	3.385	40.411	5.099	3.385	40.411	39.627	39.627
β_{23}	3.564	-3.349	73.146	74.317	74.317	54.388	6.647	6.439	55.157	54.388	54.388	54.388	4.529	3.022	46.645	45.533	1.937	0.030	40.392	1.937	0.030	40.392	39.814	39.814
β_{31}	9.221	-33.396	49.311	50.813	50.813	36.024	5.148	-10.340	36.432	36.024	36.024	36.024	3.729	-3.730	30.792	30.193	2.506	-2.236	26.552	2.506	-2.236	26.552	26.131	26.131
β_{32}	11.000	-31.336	48.967	50.785	50.785	36.117	5.752	-9.743	36.429	36.117	36.117	36.117	4.164	-3.181	30.610	29.949	2.374	-2.417	26.617	2.374	-2.417	26.617	26.123	26.123
β_{33}	12.106	-30.918	48.992	50.554	50.554	36.063	5.219	-10.396	36.449	36.063	36.063	36.063	3.951	-3.665	30.836	30.122	4.253	-0.414	26.640	4.253	-0.414	26.640	26.243	26.243

Parâmetros	N = 100						N = 500						N = 1000					
	B		B _{bnp}		EQM		B		B _{bnp}		EQM		B		B _{bnp}		EQM	
	B	B _{bnp}	B	B _{bnp}	EQM	EQM _{bnp}	B	B _{bnp}	B	B _{bnp}	EQM	EQM _{bnp}	B	B _{bnp}	B	B _{bnp}	EQM	EQM _{bnp}
β_{01}	1.118	0.031	1.534	1.477	1.477	0.602	0.195	-0.022	0.606	0.602	0.602	0.602	0.088	-0.017	0.088	-0.017	0.418	0.418
β_{02}	1.157	0.060	1.510	1.447	1.447	0.596	0.176	-0.039	0.600	0.596	0.596	0.596	0.100	-0.006	0.100	-0.006	0.417	0.417
β_{03}	1.138	0.047	1.547	1.491	1.491	0.599	0.202	-0.012	0.603	0.599	0.599	0.599	0.082	-0.024	0.082	-0.024	0.422	0.422
β_{11}	-3.101	-0.153	12.782	12.600	12.600	5.171	-0.507	0.140	5.177	5.171	5.171	5.171	-0.014	0.289	-0.014	0.289	3.604	3.609
β_{12}	-3.408	-0.374	12.568	12.374	12.374	5.184	-0.296	0.341	5.188	5.184	5.184	5.184	-0.157	0.151	-0.157	0.151	3.602	3.610
β_{13}	-3.086	-0.103	12.856	12.687	12.687	5.167	-0.522	0.116	5.171	5.167	5.167	5.167	-0.063	0.250	-0.063	0.250	3.582	3.596
β_{21}	1.727	0.292	29.380	29.029	29.029	11.953	0.113	-0.386	11.951	11.953	11.953	11.953	-0.598	-0.814	-0.598	-0.814	8.361	8.375
β_{22}	2.404	0.780	29.053	28.685	28.685	12.014	-0.430	-0.906	12.020	12.014	12.014	12.014	-0.297	-0.522	-0.297	-0.522	8.354	8.376
β_{23}	1.550	0.030	29.449	29.106	29.106	11.909	0.138	-0.348	11.910	11.909	11.909	11.909	-0.361	-0.597	-0.361	-0.597	8.302	8.329
β_{31}	1.764	-0.193	19.300	19.062	19.062	7.833	0.507	0.301	7.831	7.833	7.833	7.833	0.690	0.570	0.690	0.570	5.485	5.492
β_{32}	1.350	-0.491	19.156	18.907	18.907	7.874	0.878	0.656	7.881	7.874	7.874	7.874	0.524	0.412	0.524	0.412	5.478	5.491
β_{33}	1.915	0.013	19.287	19.041	19.041	7.786	0.493	0.279	7.787	7.786	7.786	7.786	0.512	0.406	0.512	0.406	5.449	5.462

Fonte: Produzido pelos autores.

3.4 Aplicação

Nesta seção apresenta-se uma aplicação do modelo de regressão Dirichlet a dados reais. O conjunto de dados é proveniente do *Projeto de Pesquisa Reconstrução da Estatística Pesqueira Marinha do Brasil*, financiado pelo Ministério da Pesca e Aquicultura (TED nº13/2023 - MPA/UFS), com apoio da Fundação de Apoio à Pesquisa e Extensão de Sergipe (FAPESSE). Os dados referem-se às capturas de camarão registradas no estado de São Paulo entre 1950 e 2022.

3.4.1 Imputação de dados faltantes na série histórica de capturas de camarão

O banco de dados contém as quantidades capturadas, em toneladas, das pescas artesanal e industrial, referentes às espécies camarão sete-barbas, branco, rosa e santana. Trabalha-se com a proporção de cada espécie em relação ao total capturado. Essas proporções somam 100% e, portanto, constituem uma coleção de dados composicionais.

Um exemplo do conjunto de dados para a imputação é representado na Tabela 7

Tabela 7 – Proporções anuais de captura artesanal das espécies de camarão em toneladas para imputação.

Ano	Sete-barbas	Rosa	Branco	Santana
1957	–	–	–	–
1958	–	–	–	–
1959	0.806	0.090	0.104	–
1960	0.855	0.065	0.080	–
1961	0.751	0.089	0.160	–
1962	–	–	–	–
1963	0.897	0.080	0.024	–
1964	–	–	–	–
1965	0.591	0.267	0.142	–

A imputação consiste em três etapas principais: (i) calcular as proporções das espécies nos anos com informação disponível; (ii) ajustar um modelo de regressão Dirichlet a essas proporções; e (iii) estimar as proporções para os anos com dados ausentes e convertê-las novamente em valores absolutos de captura.

Foram considerados dois cenários de aplicação:

- Imputação na série histórica da pesca artesanal das quatro espécies (sete-barbas, branco, rosa e santana), totalizando 39 observações para ajuste do modelo;
- Imputação na série histórica da pesca artesanal e industrial das espécies sete-barbas, branco e rosa, com 46 observações disponíveis para estimação.

Em ambos os casos, as proporções observadas foram modeladas em função do tempo, utilizando apenas o ano como variável explicativa. Para adequar a variável ao suporte da distribuição Dirichlet, foi aplicada a seguinte transformação:

$$Tempo_i = (I_i - I_{max}) / (I_{min} - I_{max})$$

em que $I_i = Ano_i - Ano_{init} + 1$, sendo Ano_{init} o ano do início da série (1950). O resultado da equação está entre 0 e 1.

Na primeira aplicação, ajustou-se um modelo com especificação linear para a pesca artesanal das quatro espécies:

$$\eta_{ir} = \beta_{0r} + \beta_{1r}Tempo_i + \epsilon_i, \quad i = 1, \dots, 39 \text{ e } r = 1, \dots, 4. \quad (3.27)$$

em que Y_{ir} representa o vetor de proporções das espécies no ano i . Em outras palavras, para cada ano da série histórica, organiza-se um vetor com as proporções correspondentes a cada espécie, da forma:

$$Y_i = (Y_{i,sete-barbas}, Y_{i,branco}, Y_{i,rosa}, Y_{i,santana})$$

em que cada componente indica a participação daquela espécie no total capturado no respectivo ano. A soma dessas proporções é igual a 1 (ou 100%).

As estimativas obtidas por máxima verossimilhança e suas versões corrigidas por bootstrap (paramétrico e não paramétrico) estão apresentadas na Tabela 8. Observa-se que as estimativas corrigidas diferem das estimativas originais, indicando a presença de viés nos EMV's, resultado consistente com a análise por simulação apresentada anteriormente.

As maiores discrepâncias em relação ao EMV foram observadas nos coeficientes associados à covariável, quando utilizado o bootstrap não paramétrico. Por sua vez, o bootstrap paramétrico, embora tenha apresentado diferença do EMV com menor magnitude, é recomendado para as estimativas dos interceptos.

As medidas marginais de viés e erro quadrático médio por componente, mostram estimadores com menor viés e EQM, da mesma forma como foi observado na avaliação numérica. A Tabela 9 demonstra que a precisão aumentou na versão corrigida para todas as componentes. O viés relativo também registra redução, com exceção da primeira componente (camarão sete-barbas).

A captura de camarão sete-barbas artesanal apresentou grande variação ao longo dos anos, a captura mínima foi de 38,19 toneladas (t), enquanto a máxima alcançou 2.942 t, com um percentual de 60% de dispersão em relação a tonelada média registrada ao longo da série.

Tabela 8 – Estimativas dos parâmetros pelos métodos EMV, e versões corrigidas por Bootstrap paramétrico e não paramétrico para o modelo ajustado às capturas de camarão sete-barbas, branco, rosa e santana da pesca artesanal (N = 39).

Parâmetros	EMV	EMV _{BP}	EMV _{BNP}
β_{01}	1.3709	2.6829	-0.7049
β_{02}	-0.5920	0.1395	-1.8055
β_{03}	-0.2074	0.8397	-1.1383
β_{04}	-0.2525	0.9613	0.3732
β_{11}	3.2622	3.8871	6.3985
β_{12}	2.1093	3.4448	4.0571
β_{13}	1.4224	2.1791	2.7153
β_{14}	-0.8426	-2.3919	-1.7618

Nota: BP = Bootstrap Paramétrico; BNP = Bootstrap Não Paramétrico

Fonte: Produzido pelos autores.

Isso mostra, que a técnicas de correção apresentou maior instabilidade para um cenário de alta dispersão. Também ressalte-se que tal componente apresenta a maior proporção de captura, com proporções muito próximas do limite do domínio da Distribuição Dirichlet, o que pode impactar no resultado, uma vez que proporções extremas passam por transformação, a fim de evitar erro de domínio da distribuição, uma vez que ela não está definida para os extremos do intervalo (0, 1), conforme explicado na seção 2.2.1.

A exceção observada para a componente do camarão sete-barbas artesanal pode ser explicada por característica própria da série histórica dessa categoria. Essa componente apresentou elevada variabilidade ao longo do período analisado, com captura mínima de 38,19 t e máxima de 2.942 t, correspondendo a uma dispersão aproximada de 60% em relação à tonelagem média da série.

Tabela 9 – Viés e erro quadrático médio (EQM) marginais para as componentes do modelo ajustado aos dados de capturas de camarão da pesca artesanal (N = 39).

Parâmetros	EMV		Bootstrap Paramétrico		Bootstrap Não Paramétrico	
	Viés Médio	EQM	Viés Médio	EQM	Viés Médio	EQM
$\hat{Y}_{SB,A}$	0.0097	1.0762	0.0204	0.0669	-0.0149	0.0808
$\hat{Y}_{B,A}$	-0.1378	0.8927	0.0008	0.0348	0.0105	0.0386
$\hat{Y}_{R,A}$	-0.00484	0.9538	-0.0058	0.0374	0.0016	0.0401
$\hat{Y}_{S,A}$	0.3952	2.3632	-0.0153	0.0496	0.0028	0.0513

Nota: SB = sete-barbas; B = branco; R = rosa; S = santana; A = artesanal.

Fonte: Produzido pelos autores.

Adicionalmente, o camarão sete-barbas artesanal representa a maior proporção de captura entre as componentes analisadas, assumindo frequentemente valores próximos aos limites do

domínio da distribuição Dirichlet. Proporções extremas requerem transformações prévias para evitar problemas de domínio da distribuição, conforme discutido na Seção 2.2.1. Esses fatores combinados podem auxiliar a explicar por que, diferentemente das demais componentes, não se verificou redução do viés após a aplicação das técnicas de correção.

A segunda aplicação considera um modelo com termos polinomiais, incluindo também os dados da pesca industrial. Nesse caso são considerados seis componentes (camarão sete-barbas, rosa e branco da pesca artesanal e camarão sete-barbas, rosa e branco da pesca industrial). O modelo ajustado assume a forma:

$$\begin{aligned}\eta_{i1} &= \beta_{01}, \\ \eta_{i2} &= \beta_{02} + \beta_{12} \text{Tempo}_i^2, \\ \eta_{i3} &= \beta_{03} + \beta_{13} \text{Tempo}_i^3, \\ \eta_{i4} &= \beta_{04} + \beta_{14} \text{Tempo}_i^3, \\ \eta_{i5} &= \beta_{05} + \beta_{15} \text{Tempo}_i, \\ \eta_{i6} &= \beta_{06} + \beta_{16} \text{Tempo}_i.\end{aligned}$$

As diferenças entre os estimadores na segunda aplicação estão apresentadas na Tabela 10. Assim como observado anteriormente, as estimativas corrigidas por bootstrap diferem das estimativas de máxima verossimilhança (EMV), evidenciando que a aplicação das técnicas de correção também produz alterações relevantes nos parâmetros mesmo em um modelo com especificação polinomial, estruturalmente mais complexo.

As maiores discrepâncias em relação ao EMV foram observadas no bootstrap paramétrico, que apresentou correções de maior magnitude entre os coeficientes estimados.

Os valores estimados pelo modelo apresentam menor viés e maior precisão quando usada a técnica de reamostragem bootstrap, conforme indicado na Tabela 11, em conformidade com os resultados avaliados na Simulação Monte Carlo.

Diferentemente da primeira aplicação, neste caso todas as componentes apresentaram redução das medidas de viés e erro quadrático médio após a correção. Esses resultados indicam que o desempenho dos métodos de correção não se restringe à especificação linear, estendendo-se também ao modelo polinomial. Isso reforça a efetividade das técnicas de bootstrap na regressão Dirichlet, mesmo quando o modelo incorpora maior complexidade estrutural.

3.5 Conclusão

Neste capítulo foram apresentadas abordagens para correção do viés dos estimadores na regressão Dirichlet: a abordagem baseada nas técnicas de reamostragem por bootstrap, nas versões paramétrica e não paramétrica, e a correção analítica pelo método de Cox-Snell. A avaliação do desempenho desses métodos, conduzida por meio de um estudo de simulação Monte

Tabela 10 – Estimativas dos parâmetros pelos métodos EMV, e versões corrigidas por Bootstrap paramétrico e não paramétrico para o modelo ajustado as capturas de camarão sete-barbas, branco e rosa da pesca artesanal e industrial (N = 46).

Parâmetros	EMV	EMV _{BP}	EMV _{BNP}
β_{01}	1.9304	4.2716	1.9207
β_{02}	4.3724	8.0816	4.3506
β_{03}	0.0261	1.4352	0.0260
β_{04}	4.2907	8.2227	4.2692
β_{05}	-0.1533	1.0958	-0.1526
β_{06}	2.0142	5.1788	2.0042
β_{12}	-4.0160	-6.2128	-3.9959
β_{13}	-0.5450	-1.0655	-0.5423
β_{14}	-4.7877	-7.6740	-4.7638
β_{15}	-0.1872	-0.3839	-0.1862
β_{16}	-3.8357	-7.2599	-3.8166

Nota: BP = Bootstrap Paramétrico; BNP = Bootstrap Não Paramétrico

Fonte: Produzido pelos autores.

Tabela 11 – Viés e erro quadrático médio (EQM) marginais para as componentes do modelo ajustado aos dados de capturas de camarão da pesca artesanal e industrial (N = 46).

Componentes	EMV		Bootstrap Paramétrico		Bootstrap Não Paramétrico	
	Viés Médio	EQM	Viés Médio	EQM	Viés Médio	EQM
$\hat{Y}_{SB,A}$	0.2402	1.2161	0.0290	0.0951	-0.0483	0.1356
$\hat{Y}_{SB,I}$	0.1853	1.6871	-0.0061	0.1169	-0.0165	0.1497
$\hat{Y}_{B,A}$	-0.4158	0.4967	-0.0078	0.0133	0.0204	0.0244
$\hat{Y}_{B,I}$	0.3710	0.6162	-0.0066	0.0157	0.0156	0.0208
$\hat{Y}_{R,A}$	-0.3840	0.5188	-0.0088	0.0163	0.0184	0.0246
$\hat{Y}_{R,I}$	-0.0802	1.1975	0.0003	0.0880	0.0104	0.0779

Nota: SB = sete-barbas; B = branco; R = rosa; A = artesanal; I = industrial

Fonte: Produzido pelos autores.

Carlo, permitiu analisar o comportamento do viés e do erro quadrático médio sob diferentes tamanhos amostrais e especificações de modelo.

Os resultados da avaliação numérica evidenciaram a presença de viés no estimador de máxima verossimilhança, especialmente nas amostras reduzidas ($n \leq 30$), e demonstraram que os métodos de correção foram eficazes na sua redução, impactando positivamente para ganho de precisão. De modo geral, o bootstrap paramétrico apresentou menor variabilidade entre os estimadores, exibindo mais estimadores com menor viés e menor erro quadrático médio.

A técnica de Bootstrap mostrou-se tão efetiva quanto o método analítico, que se caracteriza por ser uma metodologia com maior estabilidade na correção. Ademais, o bootstrap destaca-se por sua simplicidade de implementação.

A aplicação aos dados reais de capturas de camarão confirmou os achados da simulação, mostrando que as técnicas de correção permanecem efetivas tanto em modelos linearmente especificados quanto em modelos com termos polinomiais. Ainda que uma componente específica tenha apresentado comportamento distinto em função de elevada variabilidade e proporções extremas, os resultados gerais reforçam a utilidade das estratégias de correção propostas.

4

Seleção de Modelos Dirichlet Não-Aninhados

A inferência baseada em testes de hipóteses permite avaliar se um modelo constitui uma descrição adequada dos dados [Cribari-Neto e Lucena \(2015\)](#). Diferentemente dos critérios de informação, que buscam selecionar o modelo relativamente mais adequado dentro de um conjunto finito de alternativas, os testes de hipóteses podem levar à rejeição de todos os modelos considerados, indicando que nenhuma das especificações é capaz de descrever satisfatoriamente o processo gerador dos dados ([COX, 2013](#); [SAYYAREH, 2012](#)).

Na comparação entre modelos, pode ocorrer que um deles seja um caso particular do outro. Nesse caso, uma família de distribuições de probabilidade F é subconjunto de outra família G , denotado por $F \subset G$ quando a hipótese nula (H_0) é verdadeira, ou $G \subset F$ quando a hipótese alternativa é verdadeira. Tais modelos são denominados modelos aninhados ([COX, 2013](#)).

Por outro lado, quando nenhum modelo pode ser obtido como caso particular do outro, tem-se modelos não aninhados ([COX, 2013](#); [MCALEER, 1995](#)). Nesse cenário, as famílias F e G não apresentam correlação, e, em termos práticos, cada modelo contém ao menos um regressor que não pode ser representado como combinação linear dos regressores do modelo concorrente ([MCALEER, 1995](#)).

Nesse contexto, [Davidson e MacKinnon \(1981\)](#) propuseram o teste J para a comparação de modelos de regressão não aninhados, sob a suposição de que um dos modelos está corretamente especificado para a variável dependente y_i . O teste consiste em avaliar a hipótese nula

$$H_0 : y_i = f_i(X_i, \beta) + \epsilon_{0i} \quad (4.1)$$

contra uma hipótese alternativa

$$H_1 : y_i = g_i(Z_i, \gamma) + \epsilon_{1i} \quad (4.2)$$

em que X_i e Z_i são vetores de variáveis explicativas, β e γ são vetores de parâmetros desconhecidos, e ϵ_{0i} e ϵ_{1i} são termos de erro, usualmente assumidos como independentes e identicamente distribuídos segundo uma distribuição normal $N(0, \sigma^2)$.

Dado que H_0 e H_1 correspondem a modelos não aninhados, a evidência a favor de um implica evidência contrária ao outro. Em termos práticos, o teste J verifica se a inclusão das previsões da variável dependente obtidas a partir do modelo alternativo (H_1) no modelo sob hipótese nula (H_0) proporciona ganho significativo de poder explicativo (KELEJIAN; PIRAS, 2016). Caso contrário, não se rejeita a hipótese de que o modelo nulo está corretamente especificado. A aplicabilidade do teste e sua vasta literatura pode ser consultado em trabalhos como os de Piras e Lozano-Gracia (2012), Cribari-Neto e Lucena (2015), Cribari-Neto e Lucena (2017), Kelejian e Piras (2016).

O teste J , conforme proposto por Davidson e MacKinnon (1981), baseia-se na construção de um modelo de regressão artificial que incorpora os valores ajustados dos modelos concorrentes. Supondo $M \geq 2$ modelos de regressão não aninhados, o procedimento pode ser descrito como segue:

- Consideram-se inicialmente dois modelos não aninhados, conforme definidos em (4.1) e (4.2);
- Estima-se o modelo composto:

$$y_i = \left(1 - \sum_{j=1}^{M-1} \lambda_j\right) f_i(X_i, \beta) + \sum_{j=1}^{M-1} \lambda_j \hat{g}_{ij} + \epsilon_i, \quad (4.3)$$

em que $\hat{g}_{ij}$ representa o valor ajustado do j -ésimo modelo sob a hipótese alternativa;

- Testa-se a hipótese nula de que os coeficientes associados aos modelos alternativos são conjuntamente nulos, isto é,

$$H_0 : \lambda_1 = \lambda_2 = \dots = \lambda_{M-1} = 0, \quad (4.4)$$

utilizando-se, por exemplo, um teste t assintótico ou, o teste da razão de verossimilhança.

- Como escrito por Hagemann (2012), a estatística do teste J para testar H_1 , dado dois modelos lineares estimados por mínimos quadrados ordinários (OLS), pode ser escrito como:

$$J = \lambda_X^T \hat{\Sigma}_X^{-1} \lambda_X, \quad \text{em que} \quad (4.5)$$

$$\lambda_x = n^{1/2} \left(\mathbf{y}^\top P_Z M_x \mathbf{y} \right),$$

$$\hat{\Sigma}_x = n^{-1} \left(\mathbf{y}^\top P_Z M_x \hat{\Omega}_x M_x P_Z' \mathbf{y} \right).$$

onde $P_Z = Z(Z^\top Z)^{-1}Z^\top$ e $M_x = I - P_x$ são as matrizes de projeção de valores ajustados, e $\hat{\Omega}_x = \text{diag}(\hat{\epsilon}_{0,1}^2, \dots, \hat{\epsilon}_{0,n}^2)$ é a matriz diagonal dos vetores de resíduos quadrados $\hat{\epsilon}_{0,i} = y_i - x_i^\top \hat{\beta}$

Um aspecto relevante do teste J , destacado por Yanqin e Qi (1995), refere-se à sua sensibilidade à estrutura dos regressores nos modelos comparados. Em particular, quando os regressores do modelo sob hipótese nula são aproximadamente ortogonais aos regressores do modelo alternativo, isto é, eles são praticamente não relacionados linearmente, os valores da estatística teste J pode apresentar distorções substanciais, comprometendo a validade da inferência.

Diante disso, os autores propuseram uma versão bootstrap do teste J , visando aprimorar a aproximação da distribuição em amostras finitas da estatística do teste J . Em particular, buscando corrigir as distorções de tamanho associadas à quase ortogonalidade, bem como melhorar a aproximação baseada na teoria assintótica.

Por Yanqin e Qi (1995) o procedimento para realização do teste J com bootstrap consiste nos seguintes passos:

1. Inicialmente, usando a amostra original, estima-se o modelo sob a hipótese nula H_0 , obtendo-se o estimador de mínimos quadrados ordinários

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

bem como os resíduos estimados

$$\hat{\epsilon}_{0i} = y_i - x_i^\top \hat{\beta}, \quad i = 1, \dots, n.$$

2. Em seguida, realiza-se a amostragem com reposição dos resíduos centrados $\{\epsilon_{0i}^*\}_{i=1}^n$.

$$\hat{\epsilon}_{0i} - \bar{\epsilon}, \quad \text{em que } \bar{\epsilon} = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_{0i},$$

3. A partir desses resíduos, constrói-se uma nova variável resposta

$$y_i^* = x_i^\top \hat{\beta} + \epsilon_{0i}^*, \quad i = 1, \dots, n,$$

preservando a estrutura do modelo sob H_0 .

A amostra resultante desse novo modelo é dita ser uma amostra bootstrap.

4. Utilizando a amostra bootstrap $\{x_i, z_i, y_i^*\}_{i=1}^n$, calcula-se a estatística do teste J (ou sua forma equivalente baseada em estatísticas t).
5. Repete-se o procedimento de reamostragem e estimação um número elevado de vezes, digamos B , obtendo-se uma distribuição empírica das estatísticas do teste. Essa distribuição bootstrap é então utilizada para aproximar a distribuição amostral da estatística do teste sob H_0 .

A verificação da especificação correta requer que o teste J seja realizado seguidamente para o modelo sob a H_0 e em seguida para o modelo alternativo, como sendo o modelo correto [Cribari-Neto e Lucena \(2017\)](#). A princípio, o teste J deveria ser usado para comparar os modelos dois a dois. Para evitar essa testagem sequencial, [Hagemann \(2012\)](#) introduz o teste *minimum J* (MJ) para o caso em que múltiplos modelos não aninhados são considerados simultaneamente.

Seja um conjunto de M modelos candidatos. Para cada modelo $m : m \in M$, sob a hipótese nula, define-se a estatística do teste J_m . O teste MJ é então dado por

$$MJ = \min_{1 \leq m \leq M} J_m. \quad (4.6)$$

A validade do teste baseia-se em condições de regularidade que garantem o comportamento assintótico das estatísticas J_m , descritas em ([HAGEMANN, 2012](#)).

Se o modelo correto m^* encontra-se entre os modelos testados, isto é, se $m^* \in M$, então a estatística do teste MJ convergem, em distribuição, para χ^2_{M-1} , em caso contrário $\lim_{n \rightarrow \infty} \mathbb{P}(MJ > c) = 1$ para todo $c \in \mathbb{R}$ ([HAGEMANN, 2012](#)). O procedimento para realizar a versão bootstrap para o teste MJ , em um modelo linear, descrito por [Hagemann \(2012\)](#) pode ser realizado como:

1. Calcula-se inicialmente a menor estatística J entre os modelos comparados, e o modelo correspondente, dito $\hat{m}$;
2. Estima-se por mínimos quadrados os resíduos $\hat{\epsilon}_{0i} = y - X_{\hat{m}}\hat{\beta}_{\hat{m}}$;
3. Gera-se dados reamostrados $y^* = X_{\hat{m}}\hat{\beta}_{\hat{m}} + \epsilon_{0i}^*$;
4. Calcula-se a estatística $J^* = \lambda_{\hat{m}}^{*\top} \hat{\Sigma}_{\hat{m}}^{*-1} \lambda_{\hat{m}}^*$;
5. Calcula-se a estatística bootstrap do teste $MJ^* = \min(J^*)$;
6. Repete-se esses passos B vezes, cada repetição com nova estimação do modelo sobre a hipótese H_0 . Rejeita-se a hipótese $m^* \in M$ se a estatística MJ é maior que $c_{1-\alpha}^*$

Muito embora o procedimento acima descrito propõe o método para modelos linearmente definidos, estimados por mínimo quadrado, conforme desenvolvido por [Davidson e MacKinnon](#)

(2002), a abordagem também tem sua aplicação em modelos não lineares, como discutido em diversos estudos apresentados em [Hagemann \(2012\)](#).

Em particular, o estudo de [Cribari-Neto e Lucena \(2015\)](#) estende os testes para o contexto de modelos não lineares, considerando especificamente o modelo de regressão beta, cujos parâmetros são estimados pelo método da máxima verossimilhança. Posteriormente, [Cribari-Neto e Lucena \(2017\)](#) ampliam essa linha de pesquisa ao adaptar tais testes para uma abordagem multivariada no âmbito dos modelos de regressão GAMLSS (Generalized Additive Models for Location, Scale and Shape), classe de modelos que generaliza abordagens tradicionais ao permitir que múltiplos parâmetros da distribuição da variável resposta, como localização, escala e parâmetros relacionados à forma, sejam modelados simultaneamente como funções de covariáveis.

Nesse contexto, os testes foram estruturados para selecionar modelos não encaixados que podem diferir não apenas em seus conjuntos de regressores, mas também na especificação funcional de cada submodelo (por exemplo, na escolha das funções de ligação ou na forma paramétrica adotada para cada parâmetro da distribuição), possibilitando uma comparação ampla entre estruturas concorrentes dentro da classe GAMLSS.

4.1 Teste J e MJ para seleção de modelos Dirichlet

A regressão Dirichlet é um modelo multivariado com diferentes componentes da distribuição modelados separadamente por submodelos de regressão, permitindo que cada parâmetro, possua sua própria estrutura de covariáveis. Essa característica implica que os modelos concorrentes que seguem uma distribuição Dirichlet podem diferir no conjunto de regressores, e também na especificação de cada submodelo, o que naturalmente conduz a situações de comparação entre modelos não encaixados.

Considere o exemplo, em que há um conjunto de M modelos não aninhados candidatos para a regressão Dirichlet. O objetivo é selecionar um modelo dentre as hipóteses $H_1, H_2, \dots, H_M$. Cada hipótese H_m , com $m \in \{1, 2, \dots, M\}$, pode ser representada por

$$H_m : \begin{cases} g_{1m}(\lambda_{1m}) = \eta_{1m} = X_{1m}\beta_{1m}, \\ g_{2m}(\lambda_{2m}) = \eta_{2m} = X_{2m}\beta_{2m}, \\ \vdots \\ g_{rm}(\lambda_{rm}) = \eta_{rm} = X_{rm}\beta_{rm}, \\ \vdots \\ g_{qm}(\lambda_{qm}) = \eta_{qm} = X_{qm}\beta_{qm}. \end{cases} \quad (4.7)$$

em que q representa o número de componentes da variável resposta composicional Y .

Define-se $\delta_{\lambda_1(m)}, \delta_{\lambda_2(m)}, \dots, \delta_{\lambda_q(m)}$ como o número de regressores adicionais associados aos submodelos de cada componente r que diferem do submodelo correspondente no modelo competidor considerado. O número total de regressores adicionais utilizados no teste é dado por

$$\delta_m = \sum_{r=1}^q \delta_{\lambda_r(m)}. \quad (4.8)$$

Dessa forma, os modelos H_m e $H_{(m-1)}$ são considerados não aninhados quando ao menos um dos submodelos apresenta estrutura distinta.

O teste do modelo H_m contra um modelo competidor $H_{(m-1)}$, utilizando o teste J , pode ser realizado por meio do seguinte procedimento:

- Estimam-se inicialmente os parâmetros dos modelos competidores.
- Os preditores lineares estimados dos modelos concorrentes, denotados por $\hat{\eta}_{1i}, \hat{\eta}_{2i}, \dots, \hat{\eta}_{qi}$, para $i = 1, 2, \dots, M$ e $i \neq m$, são incluídos como regressores adicionais nos submodelos correspondentes do modelo sob a hipótese nula, H_m . O modelo resultante é denominado modelo aumentado H_a .
- Calcula-se a estatística do teste da razão de verossimilhança J , entre o modelo aumentado H_a e o modelo nulo H_m para testar a exclusão conjunta dos preditores adicionais:

$$J_m = 2(\hat{\ell}_a - \hat{\ell}_m), \quad (4.9)$$

em que $\hat{\ell}_a$ representa o valor da função de log-verossimilhança do modelo aumentado H_a e $\hat{\ell}_m$ corresponde ao valor da log-verossimilhança do modelo original H_m .

- Assim, de forma análoga ao que foi proposto por [Cribari-Neto e Lucena \(2017\)](#), a hipótese H_m é rejeitada ao nível de significância α se

$$J_m > \chi^2_{1-\alpha, \delta_m}. \quad (4.10)$$

Para ilustrar o procedimento, considere dois modelos não aninhados de regressão Dirichlet, H_1 e H_2 , com três componentes da variável resposta ($q = 3$), diferindo apenas nos submodelos associados às duas primeiras componentes η_1 e η_2 .

Para testar H_1 utilizando o teste J , procede-se da seguinte forma:

- Estimam-se os parâmetros do modelo competidor H_2 , obtendo-se os preditores lineares estimados $\hat{\eta}_{12}, \hat{\eta}_{22}$ e $\hat{\eta}_{32}$.

- Os dois primeiros preditores são incluídos como regressores adicionais nos submodelos referentes a primeiro e segundo componentes do modelo H_1 , formando o modelo aumentado. Nesse caso, o número total de regressores adicionais é $\delta_1 = 2$.
- Estima-se o modelo aumentado e calcula-se a estatística para o modelo H_1 (J_1), conforme a Equação (4.9).
- O modelo especificado em H_1 é rejeitado ao nível de significância α se

$$J_1 > \chi^2_{1-\alpha, \delta_m}. \quad (4.11)$$

O teste do modelo H_2 contra H_1 é realizado de maneira análoga, incluindo-se os preditores lineares estimados de H_1 que divergem de H_2 , em seus submodelos.

A comparação simultânea entre múltiplos modelos não aninhados pode ser realizada por meio do teste MJ . A definição dessa estatística, entretanto, requer considerações adicionais quando aplicada aos modelos de regressão Dirichlet, pois modelos concorrentes podem induzir estatísticas de teste com distribuições distintas.

A formulação original proposta por [Hagemann \(2012\)](#) fundamenta o teste MJ no princípio da Interseção e União. Segundo [Casella e Berger \(2002\)](#)[pg. 380] essa abordagem consiste em estruturar a hipótese nula global como a união de hipóteses nulas parciais associadas a cada teste que compõe o procedimento, de modo que

$$H_0 = \bigcup_{m \in M} H_{0m}$$

Para cada hipótese parcial H_{0m} considere a estatística $J_m(x)$ e sua respectiva região de rejeição R_m . Pelo princípio da interseção-união, como apresentada por [Casella e Berger \(2002\)](#) e [Lorestani e Sayyareh \(2017\)](#), se a região de rejeição de H_{0m} ocorre quando $\{x : J_m(x) \in R_m\}$, então a região de rejeição da hipótese global H_0 é dada por

$$R = \bigcap_{m=1}^M \{x : J_m(x) \in R_m\}$$

Isto significa que a região de rejeição para a hipótese global é a interseção de todas as regiões de rejeição das hipóteses individuais. Desse modo, a hipótese nula global somente é rejeitada quando todas as hipóteses parciais também são rejeitadas.

Adicionalmente, [Casella e Berger \(2002\)](#) observa que, se todas as regiões de rejeição para cada hipótese individual possuem a forma

$$R_m = \{x : J_m(x) \geq c\} \quad \forall m \in M,$$

ou seja, os testes parciais compartilham a mesma região de rejeição com valor crítico c , equivalente ao caso em que a distribuição das estatísticas $J_m(x)$ sob sua hipótese nula independe de m . Então, a região de rejeição global pode ser simplificada para

$$\bigcap_{m=1}^M \{x : J_m(x) \geq c\} = \left\{x : \inf_{m \in M} J_m(x) \geq c\right\}$$

Nessa situação, rejeita-se a hipótese global se, e somente se, a menor estatística J_m exceder o valor crítico comum, equivalente a observar valores elevados para todas as estatísticas J_m , ou, de forma equivalente, quando

$$MJ > c_{1-\alpha},$$

onde $c_{1-\alpha}$ é o quantil da distribuição assintótica do teste J .

Entretanto, no caso da regressão Dirichlet, essa simplificação descrita por [Casella e Berger \(2002\)](#) não se aplica diretamente, pois não é possível afirmar que as estatísticas J_m possuem mesma distribuição. Assim, deixa de existir, em geral, um único valor crítico comum a todos os testes parciais.

Dessa forma, a formulação baseada diretamente no mínimo $J_m(x)$ torna-se inadequada. Como alternativa, propõe-se reformular o teste global em termos dos p -valores de cada teste individual, os quais preservam a informação da distribuição específica de cada estatística J_m , e sua correspondente região de rejeição.

Assim, a regra de decisão baseando-se no princípio da Interseção-União passa a ser dada por:

$$\max(p_1, p_2, \dots, p_M) < \alpha$$

Isto significa que a hipótese nula global é rejeitada se todas as hipóteses parciais forem rejeitadas ao nível α de significância, reproduzindo a lógica da interseção das regiões de rejeição do teste MJ .

A hipótese nula do teste *Interseção-União*, é descrita como

$$H_0 : \text{existe ao menos um modelo corretamente especificado}, \quad (4.12)$$

contra a alternativa

$$H_1 : \text{todos os modelos estão incorretamente especificados}. \quad (4.13)$$

Caso a hipótese nula não seja rejeitada, seleciona-se o modelo que apresenta o maior p -valor da estatística J .

Assim como discutido por [DAVIDSON; MACKINNON](#) e [CRIBARI-NETO; LUCENA](#) também propõem a utilização do teste J com bootstrap, uma vez que, em amostras de tamanho reduzido, a distribuição da estatística do teste pode se afastar da aproximação assintótica pela distribuição qui-quadrado (χ^2).

4.1.1 Testes de seleção com Bootstrap

A versão do teste J com bootstrap, adaptada para a regressão Dirichlet, dados dois modelos não aninhados, pode ser descrito pelo Procedimento 4.1.1.

Procedimento 4.1.1 (Teste J com Bootstrap para regressão Dirichlet) *Considere dois modelos concorrentes de regressão com distribuição Dirichlet, definidos conforme a Equação 2.2. Seja H_m o modelo sob hipótese nula e H_{m-1} um modelo alternativo não aninhado. O procedimento bootstrap para implementação do teste J consiste nos seguintes passos:*

1. *Estime os modelos H_m e H_{m-1} por máxima verossimilhança e obtenha os preditores lineares*

$$\eta_{ri}(x_i; \beta_r), \quad r = 1, \dots, q, \quad i = 1, \dots, n,$$

onde q denota o número de componentes da distribuição Dirichlet e n o tamanho da amostra. Denote por $\hat{\lambda}_r$ o vetor de parâmetros estimados associados a cada componente.

2. *Construa um modelo aumentado H_a , incorporando ao modelo sob hipótese nula H_m , os preditores lineares provenientes do modelo alternativo H_{m-1} que não estejam presentes em H_m .*
3. *Para cada componente r do modelo aumentado, determine o número de preditores adicionais introduzidos em relação ao modelo nulo. Denote esse número por δ_{λ_r} . O número total de restrições testadas é dado por*

$$\delta_m = \sum_{r=1}^q \delta_{\lambda_r}.$$

4. *Calcule a estatística observada do teste J para avaliar a hipótese nula de que o modelo H_m está corretamente especificado:*

$$J = 2(\ell_a - \ell_m),$$

em que ℓ_a e ℓ_m representam, respectivamente, os valores da log-verossimilhança do modelo aumentado H_a e do modelo nulo H_m .

5. Obtenha o p -valor para a estatística de teste J , considerando δ_m graus de liberdade.
6. Sob a hipótese nula H_m , gere B amostras bootstrap por reamostragem com reposição da distribuição Dirichlet parametrizada pelos estimadores $\hat{\lambda}_r$ obtidos na estimação do modelo nulo.
7. Para cada amostra bootstrap $b = 1, \dots, B$, reestime os modelos H_m e H_a utilizando a amostra simulada.
8. Calcule a estatística bootstrap correspondente

$$J_b = 2 \times (\ell_{a,b} - \ell_{m,b}), \quad b = 1, \dots, B,$$

onde $\ell_{a,b}$ e $\ell_{m,b}$ denotam as log-verossimilhanças dos modelos aumentado e nulo estimadas na b -ésima amostra bootstrap.

9. A hipótese nula é rejeitada ao nível de significância α se J_b é maior do que o valor da estatística J observada
10. Alternativamente, o valor- p bootstrap pode ser estimado por

$$p_J^* = \frac{1}{B} \sum_{b=1}^B I(J_b > J),$$

onde $I(\cdot)$ é a função indicadora. A hipótese nula é rejeitada quando $p_J^* < \alpha$.

O procedimento para a realização do teste *Interseção-União* com bootstrap é dado por 4.1.2. O teste atua como uma regra de decisão entre os testes J individuais.

Procedimento 4.1.2 (Teste *Interseção-União* com Bootstrap para regressão Dirichlet) Considere um conjunto de M modelos candidatos de regressão Dirichlet, denotado por $\mathcal{M} = \{1, \dots, M\}$. Suponha que, para cada modelo H_m , o respectivo valor- p bootstrap tenha sido obtido conforme o Procedimento 4.1.1. O teste *Interseção-União* consiste nos seguintes passos:

1. Para cada modelo $m \in \mathcal{M}$, considere a estatística observada J_m , como descrito no procedimento 4.1.1, e obtenha o respectivo p -valor p_m associado ao teste individual.
2. Para cada replicação bootstrap $b = 1, \dots, B$, calcule as estatísticas J_b para todos os modelos $m \in \mathcal{M}$, como descrito no procedimento 4.1.1.
3. Para cada modelo, determine o p -valor da estatística bootstrap ($p_{J_m}^*$) entre os M modelos comparados, por

$$p_{J_m}^* = \frac{1}{B} \sum_{b=1}^B I(J_{bm} \geq J_m)$$

onde $I(\cdot)$ é a função indicadora.

4. *Selecione entre os modelos o p-valor superior por:*

$$p_{IU} = \max_{m \in \mathcal{M}} p_{Jm}^*.$$

5. *Ao nível de significância α , rejeita-se a hipótese nula quando $p_{IU} < \alpha$.*

6. *O modelo selecionado é aquele que apresenta o maior p-valor entre os modelos considerados:*

$$\hat{m} = \arg \max_{m \in \mathcal{M}} p_{Jm}^*.$$

7. *Em caso de rejeição da hipótese nula global, conclui-se que nenhum dos modelos considerados descreve adequadamente o processo gerador de dados.*

4.2 Avaliação Numérica

A avaliação do desempenho do teste MJ com adaptação pelo princípio da *interseção-União* no contexto da Regressão Dirichlet para amostras finitas, foi realizada por meio de simulações de Monte Carlo. Por simplicidade, são considerados $\mathcal{M} = 2$ modelos competidores com diferentes regressores.

Os valores das covariáveis foram selecionadas de uma amostragem aleatória de uma distribuição uniforme padronizada $U(0, 1)$, mantida fixa em cada réplica. As simulações envolveram 5.000 réplica de Monte Carlo, 200 reamostragem Bootstrap, e os tamanhos amostrais de tamanhos $n = 25, 50, 75$ e 100 , para avaliação foram considerado três níveis nominais: $\alpha = 1\%, 5\%$ e 10% . Os dois modelos avaliadas consideraram a seguinte estrutura linear :

$$\begin{aligned} M_1 : Y &= \begin{cases} Y_1 = \beta_1 x_1 + \beta_2 x_2 \\ Y_2 = \beta_1 x_1 + \beta_2 x_3 \end{cases} \\ M_2 : Y &= \begin{cases} Y_1 = \beta_1 x_1 + \beta_2 x_3 \\ Y_2 = \beta_1 x_1 + \beta_2 x_2 \end{cases} \end{aligned} \tag{4.14}$$

O modelo M_1 foi fixado como o modelo sob hipótese nula, isto é, ele foi considerado o modelo corretamente especificado. Enquanto o modelo M_2 , o modelo alternativo. O modelo aumentado $H_a^{(m1)}$ foi construído incluindo os preditores lineares estimados do modelo m_2 ($\eta \hat{r}_2$) no modelo m_1 , como descrito no procedimento 4.1.1, para $r = 1, \dots, q$ e $m = 1, \dots, M - 1$, como segue:

$$H_a^{(m1)} : Y = \begin{cases} Y_1 = \beta_1 x_1 + \beta_2 x_2 + \eta_{12} \\ Y_2 = \beta_1 x_1 + \beta_2 x_3 + \eta_{22} \end{cases} \quad (4.15)$$

Todas as rotinas computacionais empregadas nos estudos de simulação e nas avaliações numéricas deste capítulo foram implementadas na linguagem R (TEAM, 2024). Os experimentos foram executados na máquina com as características discutidas anteriormente na seção 3.3, em um sistema operacional Linux Ubuntu 24.04.4 LTS.

Com o objetivo de assegurar a reprodutibilidade dos resultados, bem como manter controle sobre a geração pseudoaleatória, as amostras bootstrap foram produzidas de forma sequencial, preservando uma única semente inicial ao longo de todo o experimento. Adicionalmente, visando maior eficiência computacional, especialmente diante do elevado custo associado às repetições de Monte Carlo e aos sucessivos ajustes dos modelos, foi desenvolvida uma rotina de processamento paralelo com 18 nós computacionais, para a geração das estimações entre múltiplas unidades de processamento e reduzindo o tempo total de execução das simulações.

4.2.1 Resultados do Experimento

De acordo com a Tabela 12, observa-se que a versão bootstrap do procedimento baseado no princípio da *Interseção-União* adaptada para a regressão Dirichlet apresentou taxas de rejeição da hipótese nula próximas aos níveis nominais considerados, especialmente para os níveis de significância de 1% e 5%, mostrando-se estáveis à medida que o tamanho amostral aumentou. Por exemplo, para $\alpha = 1\%$ as taxas de rejeição variaram entre 0,09% ($n = 25$) e 1,16% ($n = 100$), indicando controle do erro tipo I mesmo no menor tamanho amostral.

Para o nível de significância de 10%, as taxas de rejeição foram inferiores ao nível nominal em todos os tamanhos amostrais avaliados, indicando que o teste tendeu a rejeitar a hipótese nula com menor frequência do que o esperado pelo nível nominal. Além disso, observou-se crescimento gradual da taxa de rejeição com o aumento do tamanho amostral, aproximando-se do nível de significância considerado.

Tabela 12 – Taxas de rejeição empírica da hipótese nula (%) do procedimento bootstrap baseado no princípio da *Interseção-União*

n	$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
	25	50	75	100	25	50	75	100	25	50	75	100
p-valor $\leq \alpha$	1.06	1.02	0.90	1.16	4.12	3.46	4.12	4.20	7.52	7.34	7.84	8.34

Fonte: Produzido pelos autores.

O teste proposto nessa dissertação tem como objetivo selecionar, entre modelos concorrentes não encaixados, o modelo corretamente especificado. Tendo isso em vista, foi calculado a

proporção de vezes em que o modelo corretamente especificado foi selecionado pelo procedimento proposto. Os resultados são apresentados na Tabela 13.

Observa-se que o procedimento de seleção selecionou o modelo correto em aproximadamente 75% das simulações Monte Carlo quando $n = 25$, menor frequência observada. A proporção de seleção correta aumentou progressivamente à medida que o tamanho amostral foi ampliado. Para $n=50$, os percentuais ultrapassaram 81%. Nos cenários com $n=100$, as frequências de seleção correta superaram 92%.

Também se observa que os percentuais de seleção do modelo correto permaneceram estáveis entre os diferentes níveis de significância considerados. Esse comportamento sugere que o desempenho do procedimento mostrou-se pouco sensível à escolha do nível nominal de significância.

Tabela 13 – Frequências (%) da seleção do modelo verdadeiro utilizando o procedimento bootstrap baseado no princípio da Interseção-União como critério de seleção, quando a hipótese nula não foi rejeitada.

α	$n = 25$	$n = 50$	$n = 75$	$n = 100$
1%	74.51	80.95	87.37	92.21
5%	74.32	80.63	87.13	92.11
10%	74.05	80.53	87.02	92.10

Fonte: Produzido pelos autores.

4.3 Conclusão

Nesse capítulo apresenta-se o problema de seleção de modelos não aninhados no contexto da Regressão Dirichlet. A metodologia do teste J , e o MJ , baseada no princípio da Interseção-União, foram adaptados às particularidades da estrutura multiparamétrica da regressão Dirichlet. No caso do teste MJ , a formulação originalmente baseada na menor estatística J foi reformulada de modo que o procedimento de seleção passasse a considerar o maior p -valor associado aos testes J dos modelos comparados. Adicionalmente, foram apresentadas as versões bootstrap dos testes.

Na avaliação do desempenho, espera-se que a taxa de rejeição da hipótese nula, seja aproximadamente igual ao nível nominal considerado, de modo a controlar adequadamente o erro em decidir sobre a rejeição da hipótese nula (presença do modelo corretamente especificado entre os modelos comparados), quando ela é correta. Os resultados obtidos por meio da simulação Monte Carlo para a versão bootstrap do procedimento indicaram que as taxas de rejeição da hipótese nula aproximou-se dos níveis nominais considerados, principalmente para os níveis de significância de 1% e 5%. Para o nível de significância de 10%, observou-se aproximação progressiva ao valor nominal à medida que o tamanho amostral aumentou.

Na presença do modelo verdadeiro, o teste proposto apresentou elevada frequência de seleção correta, entre 74% (n=25) e 92% (n=100) das réplicas de Monte Carlo.

5

Considerações Finais

Os modelos de regressão Dirichlet é um instrumento para análise de dados composicionais, permitindo modelar vetores de proporções associados a diferentes componentes de um todo. Em razão de sua estrutura multiparamétrica e da flexibilidade proporcionada pela modelagem simultânea dos componentes da distribuição, esses modelos apresentam ampla aplicabilidade em diversas áreas do conhecimento.

Entretanto, a estimação dos parâmetros do modelo de regressão pelo método de máxima verossimilhança apresenta viés relevante em pequenas amostras, comprometendo a qualidade das estimativas e precisão dos resultados. Além disso, a comparação entre modelos concorrentes ainda possui limitações metodológicas na literatura voltada à regressão Dirichlet, especialmente no que se refere à adaptação de testes de hipótese originalmente formulados para modelos lineares e modelos univariados.

Diante desse cenário, esta dissertação teve como objetivo propor melhorias metodológicas para o modelo de regressão Dirichlet sob dois aspectos principais: a correção de viés dos estimadores de máxima verossimilhança por meio de técnicas bootstrap, nas versões paramétrica e não paramétrica, e a adaptação do teste MJ como critério de seleção de modelos de regressão Dirichlet não encaixados.

No estudo de correção de viés, os resultados obtidos por meio de simulações de Monte Carlo evidenciaram que os estimadores corrigidos por bootstrap apresentaram redução consistente do viés em diferentes cenários, principalmente para os menores tamanhos amostrais considerados no estudo. De modo geral, a abordagem bootstrap paramétrica apresentou desempenhos superior a versão não paramétrica, produzindo estimativas menos viesadas e com baixo impacto sobre o erro quadrático médio. Observou-se ainda que, à medida que o tamanho amostral aumenta, os estimadores originais tendem gradualmente às propriedades assintóticas esperadas, reduzindo naturalmente os efeitos do viés.

Em razão da estrutura multivariada da regressão Dirichlet e da possibilidade de diferentes

submodelos apresentarem distintos graus de liberdade sob a hipótese nula, foi proposta uma reformulação do teste MJ para basear-se nos p -valores individuais das estatísticas J , preservando o princípio da Interseção-União e permitindo considerar adequadamente as diferentes distribuições associadas aos modelos concorrentes. A adaptação proposta para o teste MJ mostrou-se adequada à estrutura do modelo composicional.

De modo geral, os resultados da simulação mostraram que o teste de seleção proposto, com o uso da técnica de bootstrap, apresentou taxas de rejeição da hipótese nula próximas aos níveis de significância estabelecidos. Isso indica que o procedimento foi capaz de controlar adequadamente o erro tipo I, isto é, a probabilidade de rejeitar incorretamente a hipótese nula quando ela é verdadeira, correspondente à decisão equivocada de concluir que o modelo corretamente especificado não está entre os modelos comparados.

Na presença do modelo verdadeiro, o teste proposto selecionou corretamente o modelo em aproximadamente 74% das réplicas para o menor tamanho amostral. O percentual de seleção aumentou progressivamente com o tamanho amostral, alcançando cerca de 92%.

Os resultados obtidos ao longo desta dissertação indicam que as metodologias propostas constituem alternativas viáveis para aprimorar a inferência em modelos de regressão Dirichlet. A utilização de técnicas bootstrap mostrou-se relevante em cenários de pequenas amostras, tanto para correção de viés quanto para melhoria do desempenho inferencial dos testes de hipótese para modelos não aninhados, comportamento consistente com a literatura sobre o método. A adaptação proposta para o teste MJ , baseada na comparação dos p -valores individuais das estatísticas J e preservando a lógica do princípio da Interseção-União, mostrou-se adequada ao contexto multiparamétrico da regressão Dirichlet.

Limitações Metodológicas

Embora os resultados obtidos nesta dissertação indiquem desempenho satisfatório das metodologias propostas, algumas limitações devem ser consideradas.

- O desempenho individual dos testes J e sua versão bootstrap não foi explorado de forma aprofundada, uma vez que o foco principal do estudo esteve voltado ao procedimento global de seleção baseado no princípio da Interseção-União e na adaptação do teste MJ ;
- Não foi realizada uma comparação direta entre a versão bootstrap do teste MJ , originalmente proposta por (HAGEMANN, 2012), e o procedimento bootstrap desenvolvido nesta dissertação. Essa comparação pode fornecer informações relevantes acerca das propriedades inferenciais de ambas as abordagens;
- O comportamento assintótico da estatística J no contexto da regressão Dirichlet não foi investigado de forma detalhada, especialmente quanto à velocidade de convergência para a

distribuição assintótica teórica e às distorções observadas em pequenas amostras;

- Os cenários de simulação considerados envolveram estruturas relativamente controladas e com número reduzido de modelos concorrentes. Dessa forma, torna-se importante avaliar o desempenho das metodologias propostas em cenários de maior complexidade, incluindo diferentes estruturas de dispersão e maior número de componentes, no processo de correção de viés;
- O presente estudo concentrou-se na avaliação do erro tipo I dos testes propostos, analisando principalmente as taxas de rejeição nula em diferentes tamanhos amostrais. Entretanto, não foi realizada uma investigação específica acerca do poder dos testes sob hipóteses alternativas. Assim, estudos futuros podem avaliar a capacidade do teste J e o MJ adaptado, e de suas versões bootstrap em identificar corretamente os modelos sob hipóteses alternativas;
- Para verificar maior precisão do teste, a ampliação da avaliação da adaptação do teste MJ com tamanho amostral e número de simulação maiores, como na literatura base (10.000 de Monte Carlo e 1.000 de bootstrap) pode ser avaliado;
- Estudos futuros podem ampliar a aplicação em dados reais, realizando comparações entre modelos concorrentes em termos de viés e erro de predição, permitindo avaliar as diferenças entre os modelos selecionados pelos testes propostos.

Referências

- AITCHISON, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 44, n. 2, p. 139–160, 1982. Citado 4 vezes nas páginas [13](#), [14](#), [18](#) e [20](#).
- AITCHISON, J. Principles of compositional data analysis. *Lecture Notes-Monograph Series*, JSTOR, p. 73–81, 1994. Citado na página [14](#).
- BOOGAART, K. G. Van den; TOLOSANA-DELGADO, R. *Analyzing compositional data with R*. [S.l.]: Springer, 2013. v. 122. Citado 2 vezes nas páginas [13](#) e [14](#).
- CASELLA, G.; BERGER, R. *Statistical inference*. 2nd edition. ed. [S.l.]: Duxbury/ Wadsworth - Thomson Learning Inc., 2002. 1–686 p. Citado 2 vezes nas páginas [63](#) e [64](#).
- CORDEIRO, G. M.; CRIBARI-NETO, F. et al. *An introduction to Bartlett correction and bias reduction*. [S.l.]: Springer, 2014. Citado 5 vezes nas páginas [15](#), [23](#), [25](#), [26](#) e [39](#).
- CORDEIRO, G. M.; KLEIN, R. Bias correction in arma models. *Statistics & Probability Letters*, Elsevier, v. 19, n. 3, p. 169–176, 1994. Citado 2 vezes nas páginas [30](#) e [32](#).
- CORDEIRO, G. M.; MCCULLAGH, P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 53, n. 3, p. 629–643, 1991. Citado 2 vezes nas páginas [24](#) e [30](#).
- COX, D. R. A return to an old paper: ‘tests of separate families of hypotheses’. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, JSTOR, p. 207–211, 2013. Citado na página [57](#).
- COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 30, n. 2, p. 248–265, 1968. Citado 5 vezes nas páginas [26](#), [27](#), [28](#), [30](#) e [31](#).
- CRIBARI-NETO, F.; LUCENA, S. E. Nonnested hypothesis testing in the class of varying dispersion beta regressions. *Journal of Applied Statistics*, Taylor & Francis, v. 42, n. 5, p. 967–985, 2015. Citado 4 vezes nas páginas [15](#), [57](#), [58](#) e [61](#).
- CRIBARI-NETO, F.; LUCENA, S. E. Non-nested hypothesis testing inference for gamlss models. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 87, n. 6, p. 1189–1205, 2017. Citado 6 vezes nas páginas [15](#), [58](#), [60](#), [61](#), [62](#) e [65](#).
- CRIBARI-NETO, F.; VASCONCELLOS, K. L. Nearly unbiased maximum likelihood estimation for the beta distribution. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 72, n. 2, p. 107–118, 2002. Citado na página [38](#).
- DAVIDSON, R.; MACKINNON, J. G. Several tests for model specification in the presence of alternative hypotheses. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 781–793, 1981. Citado 3 vezes nas páginas [15](#), [57](#) e [58](#).

- DAVIDSON, R.; MACKINNON, J. G. Bootstrap j tests of nonnested linear regression models. *Journal of Econometrics*, Elsevier, v. 109, n. 1, p. 167–193, 2002. Citado 2 vezes nas páginas [61](#) e [65](#).
- DAVISON, A. C.; HINKLEY, D. V. *Bootstrap methods and their application*. [S.l.]: Cambridge university press, 1997. Citado 4 vezes nas páginas [12](#), [37](#), [39](#) e [40](#).
- DHAENE, S.; ROSSEEL, Y. Resampling based bias correction for small sample sem. *Structural Equation Modeling: A Multidisciplinary Journal*, Taylor & Francis, v. 29, n. 5, p. 755–771, 2022. Citado na página [38](#).
- DORE, L. H. et al. Bias-corrected maximum likelihood estimation of the parameters of the complex bingham distribution. *Brazilian Journal of Probability and Statistics*, v. 30, n. 3, p. 385–400, 2016. Citado na página [38](#).
- EFRON, B. Bootstrap methods: Another look at the jackknife. *The annals of Statistics*, JSTOR, v. 7, n. 1, p. 1–26, 1979. Citado na página [37](#).
- EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: Chapman and Hall/CRC, 1994. Citado na página [37](#).
- FAHRMEIR, L. et al. *Regression: Models, methods and applications*. [S.l.]: Springer, 2013. 1–713 p. Citado 2 vezes nas páginas [12](#) e [13](#).
- FERRARI, S. L.; CRIBARI-NETO, F. On bootstrap and analytical bias corrections. *Economics Letters*, Elsevier, v. 58, n. 1, p. 7–15, 1998. Citado 3 vezes nas páginas [29](#), [30](#) e [37](#).
- FOROUTAN, H.; AMIRI, A.; KAMRANRAD, R. Improving phase i monitoring of dirichlet regression profiles. *International Journal of Reliability, Quality and Safety Engineering*, World Scientific, v. 25, n. 06, p. 1850030, 2018. Citado na página [22](#).
- GILES, D. E.; FENG, H.; GODWIN, R. T. On the bias of the maximum likelihood estimator for the two-parameter lomax distribution. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 42, n. 11, p. 1934–1950, 2013. Citado na página [38](#).
- GODWIN, R. T.; GILES, D. E. Analytic bias correction for maximum likelihood estimators when the bias function is non-constant. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 48, n. 1, p. 15–26, 2019. Citado 2 vezes nas páginas [27](#) e [28](#).
- GUEORGUIEVA, R.; ROSENHECK, R.; ZELTERMAN, D. Dirichlet component regression and its applications to psychiatric data. *Computational statistics & data analysis*, Elsevier, v. 52, n. 12, p. 5344–5355, 2008. Citado na página [22](#).
- HAGEMANN, A. A simple test for regression specification with non-nested alternatives. *Journal of econometrics*, Elsevier, v. 166, n. 2, p. 247–254, 2012. Citado 5 vezes nas páginas [58](#), [60](#), [61](#), [63](#) e [72](#).
- HIJAZI, R. H.; JERNIGAN, R. W. Modelling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics*, v. 4, n. 1, p. 77–91, 2009. Citado 5 vezes nas páginas [13](#), [18](#), [21](#), [23](#) e [24](#).
- HRON, K.; TEMPL, M.; FILZMOSER, P. Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, Elsevier, v. 54, n. 12, p. 3095–3107, 2010. Citado na página [18](#).

- KELEJIAN, H. H.; PIRAS, G. An extension of the j-test to a spatial panel data framework. *Journal of applied econometrics*, Wiley Online Library, v. 31, n. 2, p. 387–402, 2016. Citado na página 58.
- LAGOS-ÁLVAREZ, B.; JIMÉNEZ-GAMERO, M.-D.; ALBA-FERNÁNDEZ, V. Bias correction in the type i generalized logistic distribution. *Communications in Statistics—Simulation and Computation*®, Taylor & Francis, v. 40, n. 4, p. 511–531, 2011. Citado na página 38.
- LEE, H.; CHEN, S. Systematic bias of machine learning regression models and correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, USA, v. 47, n. 6, p. 4974–4983, mar. 2025. ISSN 0162-8828. Disponível em: <<https://doi.org/10.1109/TPAMI.2025.3552368>>. Citado na página 12.
- LEE, M.; WESTHOFF, P.; THOMPSON, W. The united states–china trade war and impact on the post-conservation reserve program land allocation. *Journal of Agricultural and Applied Economics*, v. 55, n. 2, p. 217–237, 2023. Citado na página 22.
- LEMONTE, A. J.; SIMAS, A. B.; CRIBARI-NETO, F. Bootstrap-based improved estimators for the two-parameter birnbaum–saunders distribution. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 78, n. 1, p. 37–49, 2008. Citado na página 38.
- LORESTANI, H.; SAYYAREH, A. Model selection using union-intersection principle for non nested models. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 46, n. 4, p. 1636–1649, 2017. Citado 2 vezes nas páginas 15 e 63.
- MACKINNON, J. G.; JR, A. A. S. Approximate bias correction in econometrics. *Journal of Econometrics*, Elsevier, v. 85, n. 2, p. 205–230, 1998. Citado na página 38.
- MAGALHÃES, T. M.; GALLARDO, D. I.; BOURGUIGNON, M. Improved point estimation for inverse gamma regression models. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 91, n. 12, p. 2444–2456, 2021. Citado 3 vezes nas páginas 27, 28 e 38.
- MAIER, M. *DirichletReg: Dirichlet Regression for Compositional Data in R*. [S.l.], 2014. (Research Report Series / Department of Statistics and Mathematics, 125). Citado 4 vezes nas páginas 18, 21, 23 e 42.
- MAVROGIORGOS, K. et al. Bias in machine learning: A literature review. *Applied Sciences*, MDPI, v. 14, n. 19, p. 8860, 2024. Citado na página 12.
- MAZUCHELI, J.; MENEZES, A. F. B.; CHAKRABORTY, S. On the one parameter unit-lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, Taylor & Francis, v. 46, n. 4, p. 700–714, 2019. Citado na página 27.
- MCALEER, M. The significance of testing empirical non-nested models. *Journal of Econometrics*, Elsevier, v. 67, n. 1, p. 149–171, 1995. Citado na página 57.
- MEDEIROS, F. M.; ARAÚJO, M. C.; BOURGUIGNON, M. Improved estimators in beta prime regression models. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 52, n. 11, p. 5125–5138, 2023. Citado na página 38.
- MELO, T. F. et al. On improved estimation in multivariate dirichlet regressions. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 49, n. 23, p. 5765–5777, 2020. Citado 5 vezes nas páginas 24, 29, 31, 35 e 37.

- MELO, T. F.; VASCONCELLOS, K. L.; LEMONTE, A. J. Some restriction tests in a new class of regression models for proportions. *Computational statistics & data analysis*, Elsevier, v. 53, n. 12, p. 3972–3979, 2009. Citado 4 vezes nas páginas 18, 20, 22 e 25.
- MENEZES, A. et al. Bias-corrected maximum likelihood estimators of the parameters of the unit-weibull distribution. *Austrian Journal of Statistics*, v. 50, n. 3, p. 41–53, 2021. Citado na página 30.
- MILLER, R. G. The jackknife-a review. *Biometrika*, Oxford University Press, v. 61, n. 1, p. 1–15, 1974. Citado na página 37.
- MORAIS, J.; THOMAS-AGNAN, C.; SIMIONI, M. Using compositional and dirichlet models for market share regression. *Journal of Applied Statistics*, Taylor & Francis, v. 45, n. 9, p. 1670–1689, 2018. Citado 2 vezes nas páginas 21 e 22.
- NARAYANAN, A. A note on parameter estimation in the multivariate beta distribution. *Computers & Mathematics with Applications*, Elsevier, v. 24, n. 10, p. 11–17, 1992. Citado 2 vezes nas páginas 22 e 24.
- ONGARO, A.; MIGLIORATI, S. A generalization of the dirichlet distribution. *Journal of Multivariate Analysis*, Elsevier, v. 114, p. 412–426, 2013. Citado na página 18.
- PALM, B. G.; BAYER, F. M.; CINTRA, R. J. Improved point estimation for the rayleigh regression model. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 19, p. 1–4, 2020. Citado 2 vezes nas páginas 27 e 28.
- PAWLOWSKY-GLAHN, V.; EGOZCUE, J. J. Compositional data and their analysis: an introduction. In: *Compositional Data Analysis in the Geosciences: From Theory to Practice*. [S.l.]: The Geological Society of London, 2006. v. 264, n. 1, p. 1–10. Citado na página 13.
- PIRAS, G.; LOZANO-GRACIA, N. Spatial j-test: some monte carlo evidence. *Statistics and Computing*, Springer, v. 22, n. 1, p. 169–183, 2012. Citado na página 58.
- SAPRA, S. K. Robust nonnested hypothesis testing. *Applied Economics Letters*, Taylor & Francis, v. 15, n. 1, p. 1–4, 2008. Citado na página 15.
- SAYYAREH, A. Inference after separated hypotheses testing: an empirical investigation for linear models. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 82, n. 9, p. 1275–1286, 2012. Citado 2 vezes nas páginas 15 e 57.
- SCHWARTZ, J.; GILES, D. E. Bias-reduced maximum likelihood estimation of the zero-inflated poisson distribution. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 45, n. 2, p. 465–478, 2016. Citado na página 26.
- SEIFOLLAHI, S.; BEVRANI, H.; ALBALAWI, O. Reducing bias in beta regression models using jackknifed liu-type estimators: Applications to chemical data. *Journal of Mathematics*, Wiley Online Library, v. 2024, n. 1, p. 6694880, 2024. Citado 2 vezes nas páginas 28 e 29.
- SHEN, Y.; YANG, Z. Bias-correction for weibull common shape estimation. *Journal of Statistical Computation and simulation*, Taylor & Francis, v. 85, n. 15, p. 3017–3046, 2015. Citado na página 26.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, Elsevier, v. 54, n. 2, p. 348–366, 2010. Citado na página 30.

SMITHSON, M.; VERKUILEN, J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, American Psychological Association, v. 11, n. 1, p. 54, 2006. Citado na página 23.

TEAM, R. C. R: *A Language and Environment for Statistical Computing*. [S.l.], 2024. Disponível em: <<https://www.R-project.org/>>. Citado 2 vezes nas páginas 42 e 68.

TÜRKAN, S.; ÖZEL, G. A jackknifed estimators for the negative binomial regression model. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 47, n. 6, p. 1845–1865, 2018. Citado 2 vezes nas páginas 28 e 29.

VASCONCELLOS, K. L.; DOURADO, G. B. Nearly unbiased estimation in a biparametric exponential family. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 78, n. 5, p. 387–404, 2008. Citado 2 vezes nas páginas 30 e 40.

WICKER, N. et al. A maximum likelihood approximation method for dirichlet's parameter estimation. *Computational statistics & data analysis*, Elsevier, v. 52, n. 3, p. 1315–1322, 2008. Citado 3 vezes nas páginas 15, 22 e 24.

YANQIN, F.; QI, L. Bootstrapping j-type tests for non-nested regression models. *Economics Letters*, Elsevier, v. 48, n. 2, p. 107–112, 1995. Citado na página 59.

Apêndices

APÊNDICE A – Implementação computacional em Linguagem R para a Correção de Viés do Modelo de Regressão Dirichlet

Nesse apêndice são apresentados o código em R para a realização da análise de simulação Monte Carlo e aplicação dos métodos propostos para a correção de viés.

A.1 Avaliação Numérica para a Correção de Viés Analítica de CoxSnell - Modelos Lineares

A.1.1 Modelos sem intercepto

```
# Pacotes necessários
library(DirichletReg)
library(fBasics)
library(pracma)
library(fpCompare)
library(compiler)

# Definição das funções para o calculo das derivadas
Jrs<-function(X,y,eta,lambda,phi){

  q<-dim(eta)[2]
  p<-dim(X)[2]
  d1Lambda<-lambda
  d2Lambda<-lambda
  digPhi<-digamma(phi)
  digLambda<-digamma(lambda)
  trigPhi<-trigamma(phi)
  trigLambda<-trigamma(lambda)

  JVal<-matrix(0,ncol=p*q,nrow=p*q)
  for(r in 1:q){
    for(a in 1:p){
```

```

j<-(r-1)*p+a
for(s in 1:q){
  for(b in 1:p){
    k<-(s-1)*p+b
    if(r==s){
      aux<-(trigPhi-trigLambda[,r])*d1Lambda[,r]^2*X[,a]*X[,b]+(digPhi-digLambda[,r]+log(y[,r]))*d2Lambda[,r]*X[,a]*X[,b]
      JVal[j,k]<-sum(aux)
    }
    else{
      aux<-trigPhi*d1Lambda[,r]*d1Lambda[,s]*X[,a]*X[,b]
      JVal[j,k]<-sum(aux)
    }
  }
}
JVal<--JVal
return(JVal)
}

Krs<-function(X,y,eta,lambda,phi){

  q<-dim(eta)[2]
  p<-dim(X)[2]
  d1Lambda<-lambda
  d2Lambda<-lambda
  digPhi<-digamma(phi)
  digLambda<-digamma(lambda)
  trigPhi<-trigamma(phi)
  trigLambda<-trigamma(lambda)

  KVal<-matrix(0,ncol=p*q,nrow=p*q)
  for(r in 1:q){
    for(a in 1:p){
      j<-(r-1)*p+a
      for(s in 1:q){
        for(b in 1:p){
          k<-(s-1)*p+b
          if(r==s){
            aux<-(trigPhi-trigLambda[,r])*d1Lambda[,r]^2*X[,a]*X[,b]
            KVal[j,k]<-sum(aux)

```

```

    }
    else{
      aux<-trigPhi*d1Lambda[,r]*d1Lambda[,s]*X[,a]*X[,b]
      KVal[j,k]<-sum(aux)
    }
  }
}
}
}

KVal<--KVal
return(KVal)
}

Krst<-function(X,y,eta,lambda,phi){

  q<-dim(eta)[2]
  p<-dim(X)[2]
  d1Lambda<-lambda
  d2Lambda<-lambda
  digPhi<-digamma(phi)
  digLambda<-digamma(lambda)
  trigPhi<-trigamma(phi)
  trigLambda<-trigamma(lambda)
  tetragLambda<-psi(2,lambda)
  tetragPhi<-psi(2,phi)

  krstVal<-array(0,c(p*q,p*q,p*q))
  for(r in 1:q){
    for(a in 1:p){
      j<-(r-1)*p+a
      for(s in 1:q){
        for(b in 1:p){
          k<-(s-1)*p+b
          for(t in 1:q){
            for(c in 1:p){
              l<-(t-1)*p+c
              if((r==s) & (r==t)){
                aux<-(tetragPhi-tetragLambda[,r])*d1Lambda[,r]^3*X[,a]*X[,b]*X[,c]+3*(
                ↪ trigPhi-trigLambda[,r])*d1Lambda[,r]*d2Lambda[,r]*X[,a]*X[,b]*X[,c]
                krstVal[j,k,l]<-sum(aux)
              }
            }
          }
        }
      }
    }
  }
}

```

```

else if((r==s) & (r!=t)){
  aux<-tetragPhi*d1Lambda[,t]*d1Lambda[,r]^2*X[,a]*X[,b]*X[,c]+trigPhi*d_j
  ↪ 1Lambda[,t]*d2Lambda[,r]*X[,a]*X[,b]*X[,c]
  krstVal[j,k,l]<-sum(aux)
}
else if((r==t) & (r!=s)){
  aux<-tetragPhi*d1Lambda[,s]*d1Lambda[,r]^2*X[,a]*X[,b]*X[,c]+trigPhi*d_j
  ↪ 1Lambda[,s]*d2Lambda[,r]*X[,a]*X[,b]*X[,c]
  krstVal[j,k,l]<-sum(aux)
}
else if((s==t) & (s!=r)){
  aux<-tetragPhi*d1Lambda[,r]*d1Lambda[,s]^2*X[,a]*X[,b]*X[,c]+trigPhi*d_j
  ↪ 1Lambda[,r]*d2Lambda[,s]*X[,a]*X[,b]*X[,c]
  krstVal[j,k,l]<-sum(aux)
}
else if((r!=s) & (r!=t) & (s!=t)){
  aux<-tetragPhi*d1Lambda[,r]*d1Lambda[,s]*d1Lambda[,t]*X[,a]*X[,b]*X[,c]
  krstVal[j,k,l]<-sum(aux)
}
}
}
}
}
}
}
return(krstVal)
}

```

```

Krs_t<-function(X,y,eta,lambda,phi){

```

```

  q<-dim(eta)[2]
  p<-dim(X)[2]
  d1Lambda<-lambda
  d2Lambda<-lambda
  digPhi<-digamma(phi)
  digLambda<-digamma(lambda)
  trigPhi<-trigamma(phi)
  trigLambda<-trigamma(lambda)
  tetragLambda<-psi(2,lambda)
  tetragPhi<-psi(2,phi)

  krstVal<-array(0,c(p*q,p*q,p*q))

```

```

    for(r in 1:q){
      for(a in 1:p){
        j<-(r-1)*p+a
        for(s in 1:q){
          for(b in 1:p){
            k<-(s-1)*q+b
            for(t in 1:q){
              for(c in 1:p){
                l<-(t-1)*q+c
                if((r==s) & (r==t)){
                  aux<-(tetragPhi-tetragLambda[,r])*d1Lambda[,r]^3*X[,a]*X[,b]*X[,c]+
                    ↪ 2*(trigPhi-trigLambda[,r])*d1Lambda[,r]*d2Lambda[,r]*X[,a]*X[,b]
                    ↪ )*X[,c]
                  krstVal[j,k,l]<-sum(aux)
                }
              else if((r==s) & (r!=t)){
                aux<-tetragPhi*d1Lambda[,t]*d1Lambda[,r]^2*X[,a]*X[,b]*X[,c]
                krstVal[j,k,l]<-sum(aux)
              }
            else if((r==t) & (r!=s)){
              aux<-tetragPhi*d1Lambda[,s]*d1Lambda[,r]^2*X[,a]*X[,b]*X[,c]+trigPhi*d
                ↪ 1Lambda[,s]*d2Lambda[,r]*X[,a]*X[,b]*X[,c]
              krstVal[j,k,l]<-sum(aux)
            }
          else if((s==t) & (s!=r)){
            aux<-tetragPhi*d1Lambda[,r]*d1Lambda[,s]^2*X[,a]*X[,b]*X[,c]+trigPh
              ↪ i*d1Lambda[,r]*d2Lambda[,s]*X[,a]*X[,b]*X[,c]
            krstVal[j,k,l]<-sum(aux)
          }
        else if((r!=s) & (r!=t) & (s!=t)){
          aux<-tetragPhi*d1Lambda[,r]*d1Lambda[,s]*d1Lambda[,t]*X[,a]*X[
            ↪ ,b]*X[,c]
          krstVal[j,k,l]<-sum(aux)
        }
      }
    }
  }
}

return(krstVal)
}

```

Função para aplicação da formula de CoxSnell

```
coxSnellBias<-function(X,y,eta,lambda,phi){

  q<-dim(eta)[2]
  p<-dim(X)[2]

  krs<-Krs(X,y,eta,lambda,phi)
  krst<-Krst(X,y,eta,lambda,phi)
  krs_t<-Krs_t(X,y,eta,lambda,phi)
  krsInv<-solve(krs)

  a<-matrix(0,ncol=p*q,nrow=p*q)
  A<-NULL
  for(t in 1:q){
    for(c in 1:p){
      l<-(t-1)*q+c
      aux<-krs_t[,l]-0.5*krst[,l]
      A<-cbind(A,aux)
    }
  }

  krsInvVec<-as.vector(krsInv)
  b<-krsInv%*%A%*%krsInvVec
  b<-matrix(b,ncol=q)
  return(b)
}
```

set.seed(381) # semente

Definição das constantes

```
RMC<-10000 # #número de replicas de Monte Carlo
n<-20 #Tamanho da amostral escolhido (20, 30, 40, 50, 60, 100, 500, 1000)
q<-3 #Número de componentes da variável resposta
p<-3 #Número de variáveis preditoras
```

Definição dos vetores de resultados

```
beta<-matrix(1,ncol=p,nrow=k)
bias<-vector("list",RMC)
correcBias<-vector("list",RMC)
```

```

set.seed(381)
# Início da Simulação Monte Carlo
enableJit(3)
for(i in 1:RMC){

  X<-matrix(runif(n*p),ncol=p)
  colnames(X)<-paste0("x",1:p)

  beta<-matrix(1,ncol=q,nrow=p)
  eta<-X%*%beta
  lambda<-exp(eta)
  phi<-apply(lambda,1,sum)

  y<-t(apply(lambda,1,rdirichlet,n=1))
  colnames(y)<-paste0("y",1:q)

  dados<-X
  dados<-as.data.frame(dados)
  dados$Y<-DR_data(y)

  aux<-paste0("x",1:p,collapse="+")
  aux<-paste0(aux,"-1")
  fitCommand<-paste0("aj<-DirichReg(Y~",aux,"",data=dados,"",model="\comm
↪ on\")")
  eval(parse(text=fitCommand))

  betaEst<-matrix(unlist(coefficients(aj)),nrow=p)
  etaEst<-X%*%betaEst
  lambdaEst<-exp(etaEst)
  phiEst<-apply(lambdaEst,1,sum)

  bias[[i]]<-betaEst-beta
  B<-coxSnellBias(X,y,etaEst,lambdaEst,phiEst)
  correcBeta<-betaEst-B
  correcBias[[i]]<-correcBeta-beta
}

# Resultados da avaliação
meanBias<-0
eqm<-0
meanCorrecBias<-0
correcEqm<-0

```

```
for(i in 1:RMC){
  meanBias<-meanBias+bias[[i]]
  meanCorrecBias<-meanCorrecBias+correcBias[[i]]

  eqm<-eqm+bias[[i]]^2
  correcEqm<-correcEqm+correcBias[[i]]^2
}

meanBias<-meanBias/RMC
meanCorrecBias<-meanCorrecBias/RMC
eqm<-eqm/RMC
correcEqm<-correcEqm/RMC
```

A.1.2 Modelos com intercepto

Na estimação do modelo com intercepto as funções e a maior parte do código mantém-se inalterados em comparação a subseção A.1.1, sendo modificada somente o seguinte trecho do código:

```
# Inicio da Simulação Monte Carlo
enableJit(3)
for(i in 1:RMC){

  # Gera a matriz de regressores
  X<-matrix(runif(n*p),ncol=p)
  X<-cbind(1,X)
  colnames(X)<-paste0("x",0:p)

  beta<-matrix(1,ncol=q,nrow=p+1)
  eta<-X%*%beta
  lambda<-exp(eta)
  phi<-apply(lambda,1,sum)

  y<-t(apply(lambda,1,rdirichlet,n=1))
  colnames(y)<-paste0("y",1:p)

  dados<-X
  dados<-as.data.frame(dados)
  dados$Y<-DR_data(y)

  aux<-paste0("x",1:p,collapse="+")
```

```

fitCommand<-paste0("aj<-DirichReg(Y~",aux,"",data=dados,"","model=\"comm
↪ on\")")
eval(parse(text=fitCommand))

betaEst<-matrix(unlist(coefficients(aj)),nrow=p+1)
etaEst<-X%*%betaEst
lambdaEst<-exp(etaEst)
phiEst<-apply(lambdaEst,1,sum)

bias[[i]]<-betaEst-beta
B<-coxSnellBias(X,y,etaEst,lambdaEst,phiEst)
correcBeta<-betaEst-B
correcBias[[i]]<-correcBeta-beta
}

```

A.2 Avaliação Numérica por Bootstrap para a Correção de Viés - Modelos Lineares

```

# Pacotes necessários
library(DirichletReg)
library(tcltk)
library(dplyr)
library(compiler)
library(xtable)

# Definição de constantes
RepMC<- 5000    #número de replicas de Monte Carlo
RepBoot<- 200   #número de amostras Bootstrap
n.comp<- 3      #número de componentes do modelo
n.vars<- 3      #número de variáveis do modelo
nSample<- 1000  #Tamanho amostral

# Definição da matriz de regressores
set.seed(381)   #semente
X<-matrix(runif(nSample*n.vars),ncol=n.vars)
X<-cbind(1,X)
colnames(X)<-paste0("x",0:n.vars)
dados<-X[,-1]
dados<-as.data.frame(dados)

# Parâmetros do modelo para inicialização

```

```

beta<- matrix(1, ncol = n.comp, nrow = n.vars+1)  #Matriz de Parâmetros
↪ verdadeiros
eta<- X%*%beta      #Usado para obter os valores estimados
lambda<-exp(eta)    #Inversa da ligação logaritmica

# Vetores para armazenar resultados
bias<-vector("list",RepMC)      #Vetor de viés
correcBias<-vector("list",RepMC) #Vetor de viés corrigido

# Início da Simulação Monte Carlo
enableJIT(3)
for(i in 1:RepMC){

  y<-t(apply(lambda,1,rdirichlet,n=1)) #Gera matriz de amostras
  ↪ aleatórias da distribuicao Dirichlet
  colnames(y)<-paste0("y",1:n.vars)
  dados$Y<-DR_data(y) #Prepara a matriz das componentes

  aux<-paste0("x",1:n.vars,collapse="+")
  fitCommand<-paste0("aj<-DirichReg(Y~",aux,",data=dados,", "model=_j
  ↪ \"common\")") #Construindo o
  ↪ modelo
  eval(parse(text=fitCommand)) #Estimando o modelo aj

  betaEst<-matrix(unlist(coefficients(aj)),nrow=n.vars+1)

  #Parâmetros do modelo ajustado
  etaEst<-X%*%betaEst
  lambdaEst<-exp(etaEst)

  betaBoot<-matrix(0,ncol=n.comp,nrow=n.vars+1)

  #Início do laço da reamostragem Bootstrap
  for(j in 1:RepBoot){

    # Realização do Bootstrap Parametrico
    yBoot<-t(apply(lambdaEst,1,rdirichlet,n=1)) #Gera matriz
    ↪ de amostras aleatorias da distribuição Dirichlet para
    ↪ cada iteracao bootstrap
    colnames(yBoot)<-paste0("y",1:n.comp)
    dadosBoot<-X[,-1]
    dadosBoot<-as.data.frame(dadosBoot)
  }
}

```

```

        dadosBoot$Y<-DR_data(yBoot)  #Matriz de componentes para cada
        ↪ iteracao bootstrap

        # Realização do Bootstrap Não Paramétrico (descomentado quando
        ↪ necessário)
        #I<- sample(1:nSample, replace = TRUE) #Amostra aleatória com
        ↪ reposição
        #dadosBoot<-X[I, -1]
        #dadosBoot<- as.data.frame(dadosBoot)
        #dadosBoot$Y<- DR_data(y[I,])  #Matriz de componentes para cada
        ↪ iteração bootstrap

        aux<-paste0("x",1:n.vars,collapse="+")
        fitCommand<-paste0("ajBoot<-DirichReg(Y~",aux,"",data=dadosBoot,
        ↪ ", "model=\"common\")") #Construindo o modelo para cada
        ↪ iteração bootstrap
        eval(parse(text=fitCommand)) #Estimando o modelo aj

        betaBoot<-betaBoot+matrix(unlist(coefficients(ajBoot)),nrow=n.vars,
        ↪ +1) } #Parâmetros do modelo ajustado em cada amostra
        ↪ bootstrap
        betaBoot<-betaBoot/RepBoot #Parâmetros relativos estimados por
        ↪ bootstrap
        bias[[i]]<-betaEst-beta #Viés dos parâmetros do modelo ajustado
        bootCorrecBeta<-2*betaEst-betaBoot #Parâmetros corrigidos por
        ↪ bootstrap
        correcBias[[i]]<-bootCorrecBeta-beta #Viés dos parâmetros do
        ↪ modelo estimado por bootstrap
    }

```

A.3 Avaliação Numérica por Bootstrap para a Correção de Viés - Modelos Polinomiais

```

# Pacotes necessários
library(DirichletReg)

# Definição de constantes
RepMC<-5000 # número de réplicas de Monte Carlo
RepBoot<-200 # número de reamostragem bootstrap
nSample<-30 # Tamanho amostral escolhido

```

```
n.comp<-3      # número de componentes do modelo
n.vars<-3      # número de variáveis resposta

# Definição da matriz de regressores
set.seed(381)  #semente
x1<-(1:(nSample+1))/(nSample+1)
x1<-x1[-(nSample+1)]
X<-x1

# Definindo covariáveis com termos polinomiais
if(n.vars>1){
  for(j in 2:n.vars){
    aux<-paste0("x",j,"<-x1^",j,";", "X<-cbind(X,x",j,")")
    eval(parse(text=aux))
  }
}
X<-cbind(1,X)
colnames(X)<-paste0("x",0:n.vars)

# Parâmetros do modelo para inicialização
aux<-c(0,1/(1:n.vars))
beta<-matrix(aux,ncol= n.comp,nrow=n.vars+1,byrow=FALSE) #Matriz de Parâmetros
↪ verdadeiros

eta<-X%*%beta
lambda<-exp(eta)
mu<-lambda/apply(lambda,1,sum) #valores estimados

# Vetores para armazenar resultados

bias<-vector("list",RepMC)      #Vetor de viés
correcBias<-vector("list",RepMC) #Vetor de viés corrigido

biasMuEst<-vector("list",RepMC)  #Vetor de vies médio
biasMuCorrecBoot<-vector("list",RepMC) #Vetor de viés médio corrigido

#Início da Simulação Monte Carlo
enableJIT(3)
for(i in 1:RepMC){

  y<-t(apply(lambda,1,rdirichlet,n=1)) ##Gera matriz de amostras
  ↪ aleatórias da distribuicao Dirichlet
```

```

colnames(y)<-paste0("y",1:n.comp)

dados<-matrix(X[,-1],ncol=n.vars)
colnames(dados)<-paste0("x",1:n.vars)
dados<-as.data.frame(dados)
dados$Y<-DR_data(y)

aux<-paste0("x",1:n.vars,collapse="+")
fitCommand<-paste0("aj<-DirichReg(Y~",aux,"",data=dados,"","model=\"comm_
↪ on\")") #Construindo o
↪ modelo
eval(parse(text=fitCommand)) #Estimando o modelo aj

betaEst<-matrix(unlist(coefficients(aj)),nrow=n.vars+1) #Parâmetros do
↪ modelo ajustado
etaEst<-X%*%betaEst
lambdaEst<-exp(etaEst)

#Parâmetros relativo para cada componente
muEst<-lambdaEst/apply(lambdaEst,1,sum)
biasMuEst[[i]]<-mu-muEst

betaBoot<-matrix(0,ncol= n.comp,nrow=n.vars+1)

#Início do laço da reamostragem Bootstrap
for(j in 1:RepBoot){

  #Realização Bootstrap Paramétrico

  yBoot<-t(apply(lambdaEst,1,rdirichlet,n=1))
  colnames(yBoot)<-paste0("y",1:n.comp)
  dadosBoot<-matrix(X[,-1],ncol=n.vars)
  dadosBoot<-as.data.frame(dadosBoot)
  dadosBoot$Y<-DR_data(yBoot)

  #Realização Bootstrap Não Paramétrico (descomentar quando
  ↪ necessário)

  #I<-sample(1:nSample,replace=TRUE)
  #dadosBoot<-matrix(X[I,-1],ncol=n.vars)
  #colnames(dadosBoot)<-paste0("x",1:n.vars)
  #dadosBoot<-as.data.frame(dadosBoot)

```

```

#dadosBoot$Y<-DR_data(y[I,])

aux<-paste0("x",1:n.vars,collapse="+")
fitCommand<-paste0("ajBoot<-DirichReg(Y~",aux,"",data=dadosBoot_
↪ ",", "model=\"common\"")
eval(parse(text=fitCommand))

betaBoot<-betaBoot+matrix(unlist(coefficients(ajBoot)),nrow=n._
↪ vars+1)
}

betaBoot<-betaBoot/RepBoot #Parâmetros do modelo ajustado em cada
↪ amostra bootstrap
bias[[i]]<-betaEst-beta #Viés dos parâmetros do modelo
bootCorrecBeta<-2*betaEst-betaBoot #Parâmetros bootstrap corrigidos
correcBias[[i]]<-bootCorrecBeta-beta #Viés dos parâmetros do modelo
↪ estimado por bootstrap

etaCorrecBoot<-X%%bootCorrecBeta
lambdaCoorecBoot<-exp(etaCorrecBoot) #Valores estimados por bootstrap
↪ para as componentes
muCorrecBoot<-lambdaCoorecBoot/apply(lambdaCoorecBoot,1,sum) #Valores
↪ relativos estimados por bootstrap para as componentes
biasMuCorrecBoot[[i]]<-mu-muCorrecBoot #Viés dos valores estimados do
↪ modelo bootstrap
}

```

A.4 Correção de viés em Modelos Dirichlet para a Captura de Pesca de Camarão

```

# Desagregar os anos de 1978 até 1986 e 1988 e 1989 do total de Camarão
↪ Artesanal em Camarao_sete_barbas_art, Camarao_rosa_art, Camarao_branco_art,
↪ Camarao_santana_art

### Definição das componentes Y
dados0<- dados

spec <-
↪ c("Cam_sete_barbas_art","Cam_rosa_art","Cam_branco_art","Cam_santana_art")
camaroes_art_total <- rowSums(dados[, spec])

```

```
# proporções das espécies
prop_camaroes <- dados[, spec] / camaroes_art_total
colnames(prop_camaroes) <- c("propCam_sete_barbas_art", "propCam_rosa_art", "propCam_branco_art", "propCam_santana_art")

dados <- cbind(dados, prop_camaroes, camaroes_art_total)

# Definição da Matriz de Regressor

tempo <- dados[, "Ano"] - 1950 + 1
tempo <- (tempo - min(tempo)) / (max(tempo) - min(tempo))
dados <- cbind(dados, tempo)
dados <- dados[, c("Ano", "tempo",
                  "propCam_sete_barbas_art",
                  "propCam_rosa_art",
                  "propCam_branco_art",
                  "propCam_santana_art",
                  "camaroes_art_total")]
dados <- na.exclude(dados)

X <- cbind(1, dados$tempo)
colnames(X) <- c("x0", "tempo")

# Formatação das proporções como Composicional
rownames(dados) <- NULL
yName <- c("propCam_sete_barbas_art", "propCam_branco_art",
          "propCam_rosa_art", "propCam_santana_art")
dados$Y <- DR_data(dados[, yName])

# Definição do modelo

bestModel <- DirichReg(Y ~ tempo | tempo | tempo | tempo, dados)

# Definição das Constantes
n.comp <- length(bestModel$n.vars) # número de componentes
n.vars <- 1                        # número de variáveis explicativas
RepBoot <- 200                     # número de reamostragem bootstrap

# Vetores para armazenar resultados
betaBoot <- matrix(0, ncol = n.comp, nrow = n.vars + 1)
lambdaBoot <- matrix(0, nrow = dim(dados)[1], ncol = n.comp)
```

```
# Definição dos parâmetros para inicialização
betaEst<- matrix(unlist(coefficients(bestModel)), nrow = n.vars+1)
lambdaEst<-predict(bestModel)

# Início do laço da reamostragem Bootstrap

dadosBoot<-dados[,2:6] #dados para a reamostragem (proporções e tempo)
dadosBoot<-as.data.frame(dadosBoot)

enableJIT(3)
set.seed(381)
for(j in 1:RepBoot){

  #--- Bootstrao Parametrico

  yBoot<-t(apply(lambdaEst,1,rdirichlet,n=1))
  colnames(yBoot)<-paste0("y",1:n.comp)
  dadosBoot$Y<-DR_data(yBoot)

  #--- Bootstrap Nao Parametrico (descomentado quando necessário)
  #I<- sample(1:dim(dadosBoot)[1], replace = TRUE)
  #dadosBoot<- dados[I,2:6]
  #dadosBoot$Y<-DR_data(dadosBoot[I,2:5])

  ajBoot<-DirichReg(formula(bestModel),data=dadosBoot,model="common")
  betaBoot<-betaEst+matrix(unlist(coefficients(ajBoot)),nrow=n.vars+1)

}

# Resultados Estimados por Bootstrap
betaBoot<-betaEst/RepBoot #Parâmetros estimados por bootstrap
bootCorrecBeta<-2*betaEst-betaBoot

lambdaBoot<- exp(X%*%bootCorrecBeta) #preditores lineares
propBoot<-lambdaBoot/apply(lambdaBoot,1,sum) #preditores em proporção

# Valores Preditores pelo modelo
propPred<- predict(bestModel)
catchPred<-propPred*dados[, "camaroes_art_total"] #Captura de camarao
↪ prevista pelo modelo
```

```

catchPredboot<-propBoot*dados[, "camaroes_art_total"] #Captura de camarao
↳ prevista pelo modelo com Bootstrap
colnames(catchPredboot)<- colnames(catchPred)

# Valores ajustados pelo modelo

I <- which(dados0$Ano %in% dados$Ano) # índices das observações ajustadas
# nomes das proporções
prop_spec <- c("propCam_sete_barbas_art", "propCam_rosa_art", "propCam_branco_ar",
↳ t", "propCam_santana_art")

# Conjunto de dados para os valores ajustados
dadoscatch <- dados0[, c("Ano", spec)]

for(i in seq_along(spec)){

  # nome das variáveis ajustadas
  est_camarao <- paste0(spec[i], "_est")
  boot_camarao <- paste0(spec[i], "_est_boot")

  # estimativas do modelo
  dadoscatch[I, est_camarao] <- catchPred[, prop_spec[i]]

  # estimativas bootstrap
  dadoscatch[I, boot_camarao] <- catchPredboot[, prop_spec[i]]
}

# valores imputados pelo modelo s/bootstrap

I<-which(is.na(dadoscatch[, "Cam_sete_barbas_art"])) # índices das
↳ observações ausentes
XFalt<-X[I,] # novos dados da variavel explicativa
impProp<-predict(bestModel, newdata=as.data.frame(XFalt)) #valores previstos
↳ para as observações ausentes
impCatch<-impProp*dados[I, "camaroes_art_total"]

# valores imputados pelo modelo c/bootstrap
impLambdaBoot<- exp(XFalt%%bootCorrecBeta) #preditores lineares para os
↳ dados ausentes
impPropBoot<-impLambdaBoot/apply(impLambdaBoot, 1, sum)
impCatchBoot<-impPropBoot*dados[I, "camaroes_art_total"]

```


APÊNDICE B – Implementação computacional em Linguagem R para a Seleção de Modelos Não-Aninhados de Regressão Dirichlet

Nesse apêndice são apresentados o código em R para a realização da análise de simulação Monte Carlo para o teste de seleção baseado no princípio da *Intersecção-União* para a Regressão Dirichlet

```
# Pacotes necessários
```

```
library(DirichletReg)
```

```
library(fBasics)
```

```
library(pracma)
```

```
library(fpCompare)
```

```
library(parallel)
```

```
# Definição de funções para cada etapa
```

```
# Função para Inicializar as variáveis e criar a formula dos modelos nulos e
```

```
↪ aumentados para realização do teste J
```

```
JTestInit<-function(A,dados){ # recebe A: matriz de regressores e matriz de
```

```
↪ dados composicionais
```

```
  M<-length(A)
```

```
  q<-dim(dados$Y)[2]
```

```
  aux1<-rep("0",q)
```

```
  modelForm0<-rep("0",M)
```

```
  fitCommand0<-rep("0",M)
```

```
  for(i in 1:M){
```

```
    for(j in 1:q){
```

```
      aux1[j]<-paste0("x",A[[i]][[j]],collapse="+")
```

```
    }
```

```
    modelForm0[i]<-paste0(aux1,collapse="|")
```

```

fitCommand0[i]<-paste0("aj<-DirichReg(Y~",modelForm0[i],"",data_
↪ =dados,"","model=\"common\"") # Cria a formula do modelo
↪ nulo
}

aux1<-rep("",q)
modelForm1<-rep("",M)
fitCommand1<-rep("",M)
delta<-rep(0,M)
for(i in 1:M){
  for(j in 1:q){
    aux1[j]<-paste0("x",A[[i]][[j]],collapse="+")
    ↪
    for(k in ((1:M)[-i])){
      if(length(setdiff(A[[k]][[j]],A[[i]][[j]]))>0){
        aux2<-paste0("eta",k,j)
        aux1[j]<-paste0(aux1[j],"+",aux2)
        delta[i]<-delta[i]+1 # Calcula o
        ↪ grau de liberdade para cada
        ↪ componente individualmente
      }
    }
  }
  modelForm1[i]<-paste0(aux1,collapse="|")
  fitCommand1[i]<-paste0("aj<-DirichReg(Y~",modelForm1[i],"",data_
↪ =dados,"","model=\"common\"") # Cria a formula do modelo
↪ aumentado
}

JTestPars<-NULL
JTestPars$modelForm0<-modelForm0
JTestPars$modelForm1<-modelForm1
JTestPars$fitCommand0<-fitCommand0
JTestPars$fitCommand1<-fitCommand1
JTestPars$delta<-delta
return(JTestPars)
}

# Função para ajustar Múltiplos modelos de regressão de Dirichlet.

fitMDirichRegs<-function(fitCommand,dados){

```

```

M<-length(fitCommand)
q<-dim(dados$Y)[2]
ll<-rep(0,M)
etaEst<-vector("list",M)
lambdaEst<-vector("list",M)
for(i in 1:M){
  eval(parse(text=fitCommand[i])) # Ajusta cada modelo criado na
  ↪ função `JTestInit`
  ll[i]<-aj$logLik # Extrai o valor da
  ↪ log-verossimilhança do modelo ajustado
  aux<-aj$fitted.values$alpha # preditores estimados
  colnames(aux)<-paste0("lambda",i,1:q)
  lambdaEst[[i]]<-aux
  aux<-log(aux)
  colnames(aux)<-paste0("eta",i,1:q)
  etaEst[[i]]<-aux
}
result<-NULL
result$ll<-ll
result$etaEst<-etaEst
result$lambdaEst<-lambdaEst

return(result)
}

```

Função para execução do test J para M modelos de regressão Dirichlet

```

JTestMModels<-function(JTestPars,resultDirichRegs,dados){
  fitCommand0<-JTestPars$fitCommand0
  fitCommand1<-JTestPars$fitCommand1
  delta<-JTestPars$delta

  M<-length(fitCommand0)
  ll1<-rep(0,M)
  ll0<-resultDirichRegs$ll
  for(i in 1:M){
    dados<-cbind(dados,resultDirichRegs$etaEst[[i]])
  }
  for(i in 1:M){
    eval(parse(text=fitCommand1[i]))
    ll1[i]<-aj$logLik

```

```

    }

    JStat<-2*(l11-l10) # Executa o teste da razão de verossimilhança entre
    ↪ o modelo nulo e o aumentado
    pvalueChisq<-1-pchisq(JStat,delta) # Extrai o pvalor baseado na
    ↪ distribuição assintótica Qui-Quadrado com delta graus de liberdade

    result<-NULL
    result$JStat<-JStat
    result$pvalueChisq<-pvalueChisq

    return(result)
}

# Função para Executar o teste J para um unico modelo.

JTest1Model<-function(model,JTestPars,resultDirichRegs,dados){

    fitCommand0<-JTestPars$fitCommand0
    fitCommand1<-JTestPars$fitCommand1[model]
    delta<-JTestPars$delta[model]

    M<-length(fitCommand0)
    l10<-resultDirichRegs$l1[model]
    for(i in 1:M){
        dados<-cbind(dados,resultDirichRegs$etaEst[[i]])
    }

    eval(parse(text=fitCommand1))
    l11<-aj$logLik

    JStat<-2*(l11-l10)
    pvalueChisq<-1-pchisq(JStat,delta)

    result<-NULL
    result$JStat<-JStat
    result$pvalueChisq<-pvalueChisq

    return(result)
}

```

Função para calcular as estatísticas dos testes J usando o bootstrap
↪ paramétrico em processamento paralelo.

```
parallelJStatBoot<-function(i,JTestPars,dadosBoot){ # dadosBoot contem todas as
  ↪ reamostragem para cada modelo
    resultDirichRegs<-fitMDirichRegs(JTestPars$fitCommand0,dadosBoot[[i]])
    M<-length(JTestPars$fitCommand0)
    j<-(i-1)%M+1 # Operacao que permite identificar de qual modelo foi
      ↪ gerado o conjunto de dados i
    JStat<-JTest1Model(j,JTestPars,resultDirichRegs,dadosBoot[[i]])$JStat
    return(JStat)
}
```

Especificação dos hiperparâmetros dos modelos

```
n<-25 #Tamanho da amostra escolhida (25, 50, 75 e 100)
q<-2 #Número de componentes da variavel resposta
p<-3 #Número máximo de variáveis preditoras
M<-2 #Número de modelos concorrentes
```

Cria lista dupla de índices

#A primeira camada indica o modelo M, a segunda indica a componente que está
↪ sendo ajustada. Os indices de cada camada define as variaveis preditoras
↪ que fazem parte do modelo.

```
A<-vector("list",M)
for(i in 1:M){
  A[[i]]<-vector("list",q)
}
```

Define os indices das variaveis preditoras do Modelo verdadeiro

```
A[[1]][[1]]<-c(1,2)
A[[1]][[2]]<-c(1,3)
```

Define os indices das variaveis preditoras do Modelo alternativo (falso)

```
A[[2]][[1]]<-c(1,3)
A[[2]][[2]]<-c(1,2)
```

Definição da matriz de regressores

```
set.seed(381) # semente
```

```
X<-matrix(runif(n*p),ncol=p)
X<-cbind(1,X)
colnames(X)<-paste0("x",0:p)
dados<-X[,-1]
dados<-as.data.frame(dados)

# Parâmetros do modelo para inicialização
beta<-matrix(0,ncol=q,nrow=p)
for(j in 1:q){
  beta[A[[1]][[j]],j]<-1
}
beta<-rbind(1,beta)

eta<-X%%beta      # Usado para obter os valores estimados
lambda<-exp(eta)  # Inversa da ligação logaritimica
phi<-apply(lambda,1,sum) # Parâmetro de concentração

y<-t(apply(lambda,1,rdirichlet,n=1)) #Gera matriz de amostras aleatorias da
↪ distribuição Dirichlet
colnames(y)<-paste0("y",1:q)
dados$Y<-DR_data(y) # Prepara a matriz de componentes

# Inicialização do teste de seleção
JTestPars<-JTestInit(A,dados)

# Definição das constantes
RBoot<-200
RMC<-5000
I<-1:(RBoot*M)

# Vetores para armazenar resultados
pvalueChisq<-matrix(0,nrow=RMC,ncol=M)
pvalueBoot<-rep(0,RMC)
JStat<-matrix(0,nrow=RMC,ncol=M)
selModel<-rep(0,RMC)
selModelBoot<-rep(0,RMC)
dadosBoot<-vector("list",RBoot*M)

# Início da Simulação Monte Carlo
set.seed(381)
```

```
for(i in 1:RMC){

  #Gerando amostras de Monte Carlo

  y<-t(apply(lambda,1,rdirichlet,n=1))
  colnames(y)<-paste0("y",1:q)

  dados<-X[,-1]
  dados<-as.data.frame(dados)
  dados$Y<-DR_data(y)

  #Calculando EMV's dos parâmetros da regressão
  resultDirichRegs<-fitMDirichRegs(JTestPars$fitCommand0,dados)
  lambdaEst<-resultDirichRegs$lambdaEst

  #Realizando o testes J
  JTestResult<-JTestMModels(JTestPars,resultDirichRegs,dados)
  JStat[i,<-JTestResult$JStat

  selModel[i]<-which.min(JTestResult$JStat) # seleciona o modelo com a menor
  ↳ estatística J
  pvalueChisq[i]<-JTestResult$pvalueChisq[selModel[i]]

  #Gerando amostras Bootstrap
  cont<-0
  for(j in 1:RBoot){
    for(k in 1:M){
      cont<-cont+1
      yBoot<-t(apply(lambdaEst[[k]],1,rdirichlet,n=1))
      dadosBoot[[cont]]<-as.data.frame(X[,-1])
      dadosBoot[[cont]]$Y<-DR_data(yBoot)
    }
  }

  #Calculando réplicas bootstrap das estatísticas J e MJ

  aux<-mclapply(I,parallelJJStatBoot,mc.cores=18,dados=dadosBoot,JTestPar
  ↳ s=JTestPars,mc.preschedule=TRUE)
  aux<-unlist(aux)
  JStatBoot<-matrix(aux,ncol=M,byrow=TRUE)

  aux<-rep(0,M) # Armazena o pvalor bootstrap
```

```
# Calculando o p-valor para estatística J bootstrap
  for(j in 1:M){
    aux[j]<-mean(JStatBoot[,j]>=JStat[i,j])
  }

# Teste de seleção proposto
  selModelBoot[i]<-which.max(aux)
  pvalueBoot[i]<-max(aux)
}
close(pb)
```