UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Natasha Ferrari Lopes

Modelando o desempenho dos campeões do Brasileirão através de regressões no intervalo unitário

Juiz de Fora
2025

Natasha Ferrari Lopes

Modelando o desempenho dos campeões do Brasileirão através de regressões no intervalo unitário

Trabalho de conclusão de curso apresentado ao Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Maia Magalhães

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Lopes, Natasha Ferrari.
Modelando o desempenho dos campeões do Brasileirão através de regressões no intervalo unitário / Natasha Ferrari Lopes. – 2025.
46 f. : il.

Orientador: Tiago Maia Magalhães
Trabalho de Conclusão de Curso (graduação) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Departamento de Estatística, 2025.

1. modelagem estatística. 2. variável unitária. 3. comparação de modelos. I. Magalhães, Tiago Maia, orient. II. Modelando o desempenho dos campeões do Brasileirão através de regressões no intervalo unitário.

Natasha Ferrari Lopes

Modelando o desempenho dos campeões do Brasileirão através de regressões no intervalo unitário

Trabalho de conclusão de curso apresentado ao Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em 18 de agosto de 2025

BANCA EXAMINADORA

Prof. Dr. Tiago Maia Magalhães - Orientador
Universidade Federal de Juiz de Fora

Profª. Drª. Bárbara da Costa Campos Dias
Universidade Federal de Juiz de Fora

Prof. Dr. Clécio da Silva Ferreira
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Agradeço a Deus, por me conceder força, sabedoria e coragem para trilhar esta jornada. À minha família, especialmente aos meus pais, João Carlos e Ana Maria, por toda a dedicação e incansável empenho para que eu chegasse até aqui. Às minhas irmãs, Tainan e Lorena, pelo constante incentivo e por tornarem mais leves os momentos difíceis. Ao meu namorado, Alysson, pelo apoio incondicional e por acreditar em mim em cada etapa desse percurso. Aos meus professores, cuja instrução foi fundamental, em especial ao meu orientador, Tiago, pela paciência e aprendizado compartilhado. Aos meus amigos e colegas de classe, pela convivência, pelas trocas enriquecedoras e amizade cultivada ao longo desta caminhada.

RESUMO

Em diversos contextos, a modelagem estatística requer a consideração de restrições impostas à variável resposta, como a limitação ao intervalo unitário em proporções. Nesses casos, é comum empregar distribuições compatíveis com essa faixa, embora modelos com suporte mais amplo também sejam passíveis de aplicação. Nesta monografia, são abordados os modelos Beta, Simplex e Normal na análise de dados entre 0 e 1. A distribuição Normal é utilizada como referência, com o propósito de verificar a paridade do seu desempenho em relação ao das distribuições formuladas para esse perfil de variável. Assim, foi discutido os três modelos de regressão e seus processos de estimação pelo Método da Máxima Verossimilhança. Para avaliar a eficácia dos modelos, realizou-se um estudo de simulação usando o método de Monte Carlo para estimar os parâmetros de regressão, observando que as estimativas se aproximam dos valores reais conforme o tamanho da amostra aumenta. Outrossim, os modelos estatísticos foram aplicados ao conjunto de dados dos clubes campeões do Brasileirão entre os anos de 2003 e 2024, utilizando o aproveitamento de pontos como variável de interesse com o propósito de demonstrar sua implementação prática e a interpretação dos resultados obtidos. Na aplicação, são empregadas técnicas de visualização, estatísticas descritivas e regressões adequadas aos dados. Os estudos de simulação e a aplicação apresentados neste trabalho confirmam os objetivos propostos e fornecem contribuições para futuras pesquisas.

Palavras-chave: modelagem estatística; intervalo unitário; simulação de Monte Carlo.

ABSTRACT

In many contexts, statistical modeling requires consideration of constraints imposed on the response variable, such as the limitation to the unitary interval for proportions. In these cases, it is common to employ distributions compatible with this range, although models with broader support are also applicable. This monograph addresses the Beta, Simplex, and Normal models in the analysis of data between 0 and 1. The Normal distribution is used as a reference to verify its performance parity with that of distributions formulated for this variable profile. Thus, the three regression models and their estimation processes using the Maximum Likelihood Method were discussed. To assess the effectiveness of the models, a simulation study was conducted using the Monte Carlo method to estimate the regression parameters, observing that the estimates approach the true values as the sample size increases. Furthermore, the statistical models were applied to the dataset of Brazilian Championship champion clubs between 2003 and 2024, using points per share as the variable of interest, to demonstrate their practical implementation and the interpretation of the results. Visualization techniques, descriptive statistics, and regressions appropriate to the data were employed in the application. The simulation studies and the application presented in this work confirm the proposed objectives and provide contributions for future research.

Keywords: statistical modeling; unitary interval; Monte Carlo simulation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

FDP	Função Densidade de Probabilidade
a.a.	Amostra Aleatória
SMC	Simulação de Monte Carlo
AIC	Critério de Informação de Akaike
BIC	Critério de Informação Bayesiano

SUMÁRIO

1 INTRODUÇÃO

Em diferentes cenários, é possível explicar o comportamento de determinadas variáveis com base em dados relacionados ao fenômeno investigado. O estudo de Galton (1889), por exemplo, aplicado no campo da Antropologia, propõe uma análise voltada a investigar se a altura dos filhos pode ser explicada a partir da altura dos pais. A verificação dessas relações pode ser estendida a demais áreas do conhecimento. Nomeadamente, a pesquisa de Trunfio et al. (2022), que buscou prever o tempo total de permanência hospitalar de pacientes submetidos à apendicectomia laparoscópica, utilizando variáveis clínicas relacionadas ao estado de saúde dos pacientes, variáveis biológicas como idade e sexo, além de fatores associados ao processo de internação, como o tempo de espera até a realização da cirurgia. Outra demonstração é o trabalho de Zaluska e Gładyszewska-Fiedoruk (2020) que investigou a relação entre a qualidade do ar e a incidência de doenças respiratórias agudas em crianças.

Nessa perspectiva, emerge a necessidade do uso de modelos estatísticos, que, por sua vez, permitem não apenas identificar e compreender relações entre variáveis dentro de um conjunto de dados, mas também realizar inferências sobre o contexto estudado, promovendo uma análise fundamentada e contribuindo significativamente para a tomada de decisões em diversas áreas do conhecimento. Um exemplo prático da aplicação de tais modelos pode ser encontrado no estudo de Kieschnick e McCoullough (2003), que analisaram o impacto da implementação de um plano de saúde baseado em evidências na produtividade e nos resultados de saúde de bezerros criados em fazendas leiteiras na Grã-Bretanha. O uso de análise estatística permitiu aos autores mensurar os efeitos da intervenção de forma rigorosa, contribuindo para a compreensão e melhoria de práticas no setor agropecuário.

Sob esse ponto de vista, adota-se o conceito de regressão, uma técnica estatística utilizada na modelagem de dados, que viabiliza quantificar a relação entre uma variável de interesse e uma ou mais variáveis explicativas. Segundo Neter et al. (1996), essa metodologia é amplamente utilizada em negócios, ciências sociais e biológicas e muitas outras disciplinas. Ademais, relações lineares entre variáveis são bastante frequentes em contextos estatísticos. Nesse tipo de associação, uma variação nas variáveis independentes implica uma variação proporcional na variável dependente. Esse comportamento pode ser representado por um modelo simples, quando há apenas uma variável explicativa, ou por um modelo múltiplo, quando envolve duas ou mais variáveis independentes.

A modelagem de dados no intervalo unitário surge como uma extensão natural das abordagens tradicionais de regressão, direcionando-se a técnicas estatísticas específicas para variáveis de interesse cujos valores estão restritos entre 0 e 1. Essa restrição é exemplificada por proporções, que nem sempre podem ser modeladas por métodos tradicionais sem

que haja violação de pressupostos ou a geração de observações fora do intervalo válido. Nesse sentido, Lima (2018) destaca a importância de utilizar modelos probabilísticos que sejam, ao mesmo tempo, realistas e suficientemente flexíveis para capturar a complexidade inerente aos dados, preservando as características fundamentais dessas variáveis.

Neste cenário, existem várias distribuições de probabilidade que podem ser consideradas para a modelagem estatística, entre elas as distribuições Beta e Simplex. Segundo Cribari-Neto e Zeileis (2010), a distribuição Beta reparametrizada (Cribari-Neto e Ferrari (2004)) revela-se uma opção atrativa para esse tipo de dado devido à sua flexibilidade, decorrente da forma como sua densidade é descrita em termos da média μ e do parâmetro de precisão ϕ, podendo assumir várias formas diferentes, dependendo da combinação dos valores dos parâmetros. A regressão beta traz consigo as vantagens de abordagens paramétricas no contexto de modelagem, como a interpretação direta dos parâmetros, menor complexidade, menor exigência de grandes amostras e menor custo computacional, entre outras. No entanto, conforme discutido por Zerbinatti e Ferrari (2008), em cenários preditivos, a função de regressão estimada pode subestimar os valores da variável resposta. Em complemento, a distribuição Simplex, proposta por Barndorff-Nielsen e Jørgensen (1991), é frequentemente utilizada como uma alternativa à Beta, por sua capacidade igualmente flexível e por apresentar robustez em pequenas amostras, devido à sua formulação baseada na variância da média amostral. Adicionalmente, de acordo com Cribari-Neto e Zeileis (2010), taxas e proporções geralmente apresentam distribuições assimétricas, o que torna as aproximações baseadas na distribuição Normal potencialmente imprecisas, visto que pressupõe simetria e suporte em toda a reta real. Apesar disso, opta-se por avaliar a distribuição Normal, devido à sua notoriedade e às vantagens que oferece - como parâmetros de fácil interpretação, formas analíticas simples e tratamento algébrico acessível. Além disso, está disponível na maioria dos softwares estatísticos e tem ampla aplicabilidade, sendo capaz de modelar fenômenos variados em diferentes áreas do conhecimento. Tais conceitos e propriedades mencionados serão discutidos no próximo capítulo.

1.1 OBJETIVO

Por conseguinte, este trabalho busca estudar distribuições aplicadas no intervalo unitário com o objetivo de investigar suas propriedades de ajuste frente a proporções modeladas como variáveis contínuas, considerando tanto abordagens tradicionalmente recomendadas quanto alternativas amplamente utilizadas em múltiplas situações. A proposta envolve a análise comparativa de três modelos estatísticos: Beta, Simplex e Normal, que se diferenciam em termos de suporte teórico e estrutura de distribuição, a fim de examinar sua adequação e desempenho na modelagem de dados restritos ao intervalo (0,1). Para tanto, são conduzidas simulações controladas e uma aplicação empírica que permite observar, em distintas condições, o comportamento dos estimadores e a qualidade

dos ajustes proporcionados por cada modelo. Dessa forma, deseja-se oferecer uma avaliação fundamentada sobre o uso dessas distribuições, contribuindo para a escolha mais apropriada de técnicas conforme as características específicas dos dados analisados.

1.2 ORGANIZAÇÃO

Mediante a exposição introdutória realizada, este trabalho está estruturado em três capítulos, cada um com um propósito específico que contribui de forma complementar para a análise proposta.

O Capítulo 2 é dedicado à apresentação dos modelos estatísticos utilizados, baseados nas distribuições Beta, Simplex e Normal, com foco na modelagem de variáveis contínuas restritas ao intervalo unitário. Nesse capítulo, são discutidas as propriedades teóricas das distribuições, suas respectivas funções de densidade, os parâmetros envolvidos e as estruturas de regressão adotadas, além dos métodos de estimação e pacotes estatísticos empregados.

O Capítulo 3 apresenta um estudo de simulação, no qual os modelos descritos previamente são avaliados em diferentes cenários, com o intuito de investigar o comportamento das estimativas obtidas, considerando diferentes tamanhos amostrais e estruturas paramétricas.

O Capítulo 4 refere-se a uma aplicação prática dos modelos em dados reais, permitindo verificar sua adequação empírica e potencial interpretativo diante de uma situação concreta. São discutidos os resultados obtidos e comparadas as performances dos modelos no contexto aplicado.

Por fim, o Capítulo 5 traz as considerações finais, destacando os principais achados da pesquisa, suas limitações e possíveis direções para estudos futuros.

2 ESPECIFICAÇÃO DOS MODELOS DE REGRESSÃO

Este capítulo tem como objetivo apresentar os modelos estatísticos baseados nas distribuições Beta, Simplex e Normal, utilizados neste trabalho para a análise de variáveis contínuas restritas ao intervalo unitário. Além de fornecer os fundamentos teóricos necessários para a compreensão dos ajustes realizados tanto nas simulações quanto na aplicação prática dos modelos. Inicialmente, será feita a descrição de cada distribuição, destacando suas propriedades, funções densidade de probabilidade e parâmetros envolvidos. Em seguida, serão especificadas as estruturas de regressão adotadas para cada modelo, com ênfase na forma como a média e a dispersão são modeladas em função de covariáveis. Também serão discutidos os pressupostos e particularidades de cada abordagem, bem como os métodos de estimação empregados, além dos pacotes estatísticos utilizados para modelagem.

No processo de estimação, destaca-se o uso do método de Newton-Raphson para a resolução das equações de máxima verossimilhança. Esse método iterativo baseia-se no cálculo sucessivo das derivadas da função de log-verossimilhança, permitindo a aproximação eficiente dos estimadores dos parâmetros. Trata-se de uma técnica amplamente utilizada em estatística por sua rapidez de convergência, desde que as condições de regularidade sejam atendidas (Akram e Qurrat (2015)).

2.1 MODELO BETA

A distribuição Beta é uma família de distribuições contínuas definida no intervalo unitário, caracterizada por dois parâmetros de forma, usualmente denotados por α e β. Sua grande flexibilidade permite representar diversas formas, desde simétricas até marcadamente assimétricas, o que a torna especialmente apropriada para modelar variáveis desta natureza. Essa versatilidade estrutural é particularmente vantajosa em contextos aplicados, nos quais os dados apresentam comportamentos variados ao longo do intervalo. Além disso, a Beta é amplamente reconhecida e utilizada em diferentes áreas do conhecimento, destacando-se como uma ferramenta estatística eficaz para a análise de proporções e taxas.

2.1.1 Caracterização Teórica do Modelo

A função densidade de probabilidade (FDP) da distribuição Beta, com parâmetros $\alpha, \beta > 0$ é dada por:

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \ 0 \leq y \leq 1,$$

em que $\Gamma(.)$ é a função gama.

O modelo de regressão Beta, proposto por Cribari-Neto e Ferrari (2004), adota uma parametrização alternativa da função densidade de probabilidade, expressa em termos

dos parâmetros de média μ e precisão ϕ. Para isso, os parâmetros usuais da distribuição Beta são reparametrizados da seguinte forma:

$$\alpha = \mu\phi, \tag{2.1}$$
$$\beta = (1-\mu)\phi. \tag{2.2}$$

Dessa forma, sejam $y_1, y_2, \ldots, y_n$ uma amostra aleatória de uma variável que segue a distribuição Beta reparametrizada, tal que $Y \sim \text{Beta}(\mu\phi, (1-\mu)\phi)$, sua FDP é definida como:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\Big[(1-\mu)\phi\Big]} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \ 0 < y < 1,$$

em que $\mu \in (0,1)$ e $\phi > 0$ com valor esperado e variância sendo:

$$\mathbb{E}(Y) = \mu,$$
$$\text{Var(Y)} = \frac{\mu(1-\mu)}{1+\phi}. \tag{2.3}$$

A estrutura da variância permite capturar a heterocedasticidade típica de variáveis restritas ao intervalo unitário, já que a variância é diretamente influenciada pela própria média. Além disso, fixado o valor de μ, quanto maior for o valor do parâmetro de precisão ϕ, menor será a variância de Y. Essa observação também pode ser verificada por meio dos gráficos da FDP da distribuição Beta, considerando diferentes valores de μ e ϕ.

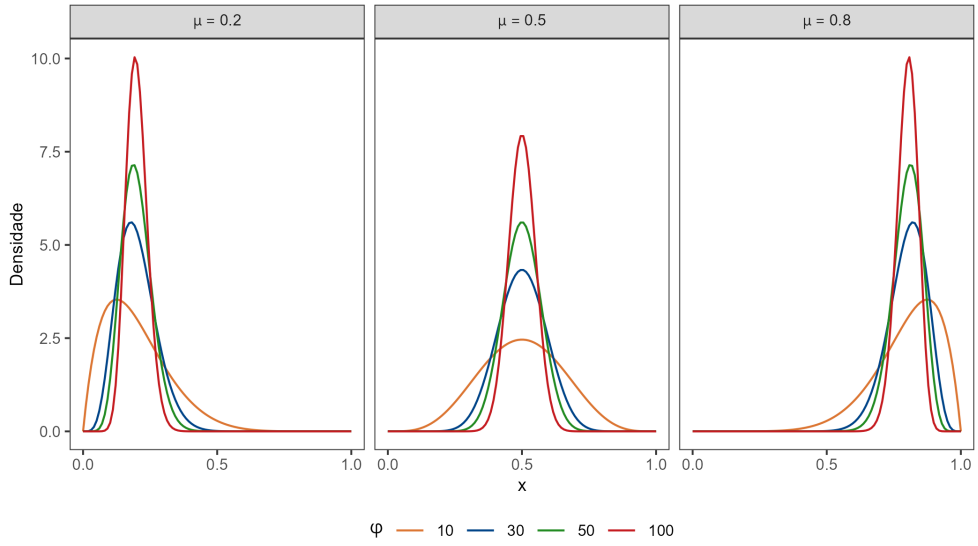

Figura 1 – Densidade da distribuição Beta para diferentes combinações de μ e ϕ.

Fonte: Elaboração do autor.

A partir da Figura 1, observa-se que, com o aumento do parâmetro ϕ, a curva se torna mais estreita e alta, indicando menor variabilidade e maior concentração dos valores próximos à média. Além disso, com base nas Equações (2.1) e (2.2), quando a média

$\mu = 0.5$, a distribuição Beta é simétrica, já que $\alpha = \beta$, resultando em um gráfico espelhado no centro. Para $\mu < 0.5$, como $\alpha < \beta$, gera assimetria à direita. Já para $\mu > 0.5$, com $\alpha > \beta$, a cauda se alonga para a esquerda.

Em sua forma mais simples, o parâmetro de precisão é assumido constante para todas as observações. No entanto, extensões do modelo permitem especificar uma subestrutura para ϕ como função de covariáveis, possibilitando modelar a variabilidade da dispersão entre as observações empregada por Smithson e Verkuilen (2006) e apresentada por Simas et al. (2010). Neste trabalho, será considerado o caso de ϕ constante.

Os seguintes passos foram baseados nos trabalhos de Cribari-Neto e Zeileis (2010), Oliveira (2004) e Andrade (2007). Posto isso, a reparametrização possibilita incorporar a média como função de covariáveis por meio de uma estrutura de regressão. Especificamente, a média é modelada como:

$$g(\mu_i) = \eta_i = x_i^\top \beta, \tag{2.4}$$

em que $g(.)$ é a função de ligação, $x_i = (x_{i1} \ldots, x_{ik})$ é o vetor de k covariáveis, $\beta = (\beta_1, \ldots, \beta_k)^\top$ é o vetor $k \times 1$ de coeficientes associados às covariáveis ($k < n$) e η_i é um preditor linear (ou seja, $\eta_i = \beta_1 x_{i1}, \ldots, \beta_k x_{ik}$; geralmente $x_{i1} = 1$ para todo i, de modo que o modelo tenha um intercepto).

Para garantir que a média μ_i permaneça no intervalo (0,1), adotamos a função de ligação logito, que é definida por:

$$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right). \tag{2.5}$$

Invertendo essa relação, e considerando a definição de η_i na Equação (2.4), obtém-se a expressão explícita para a média em função das covariáveis:

$$\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}.$$

A partir dessa parametrização, a variância da variável resposta pode ser expressa conforme a Equação (2.3), resultando em:

$$\text{Var(Y)} = \frac{g^{-1}(x_i^\top \beta)[1 - g^{-1}(x_i^\top \beta)]}{1+\phi}.$$

Para estimar os parâmetros β e ϕ, utiliza-se o método da máxima verossimilhança. A função de log-verossimilhança total é dada pela soma das contribuições individuais das observações, conforme

$$l(\beta, \phi) = \sum_i l_i(\mu_i, \phi),$$

sendo que cada contribuição $l_i(\mu_i, \phi)$ tem a forma:

$$\begin{aligned} l_i(\mu_i, \phi) = \; & \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma[(1-\mu_i)\phi] \\ & + (\mu_i\phi - 1)\log y_i + [(1-\mu_i)\phi - 1]\log(1-y_i). \end{aligned} \tag{2.6}$$

Ao substituir μ_i pela sua parametrização em função das covariáveis, conforme a Equação (2.4), a função de log-verossimilhança torna-se dependente diretamente dos parâmetros β e ϕ.

Definem-se $y_i^* = \log\{y_i/(1-y_i)\}$ e $\mu_i^* = \psi(\mu_i\phi) - \psi((1-\mu_i)\phi)$, em que $\psi(.)$ é a função digama. A diferenciação da função de log-verossimilhança com respeito aos parâmetros desconhecidos (β,ϕ), definida como função escore, é dada por:

$$\frac{\partial}{\partial\beta} = \phi X^\top \mathrm{T}(y^* - \mu^*), \tag{2.7}$$

$$\frac{\partial}{\partial\phi} = \sum_{i=1}^{n} \left\{\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi((1-\mu_i)\phi) + \psi(\phi)\right\}, \tag{2.8}$$

no qual X é uma matriz $n \times k$, sendo $y_i^\top$ a i-ésima linha dessa matriz, $y^* = (y_1^*, \ldots, y_n^*)^\top$, $T = diag\{1/g'(\mu_i), \ldots, 1/g'(\mu_n)\}$ e $\mu^* = (\mu_1^*, \ldots, \mu_n^*)^\top$.

A matriz de informação de Fisher, correspondente à segunda derivada da log-verossimilhança (tomada com sinal negativo), é utilizada tanto para auxiliar no processo iterativo do algoritmo quanto para derivar as propriedades assintóticas dos estimadores. Essa matriz pode ser particionada da seguinte forma:

$$K = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix},$$

em que

$$K_{\beta\beta} = \phi^2 X^\top W X, \quad K_{\beta\phi} = K_{\phi\beta}^\top = X^\top T c, \quad K_{\phi\phi} = \mathbf{1}^\top D \mathbf{1}.$$

Nessa expressão, $W = \mathrm{diag} w_1, \ldots, w_n$, com

$$w_i = \phi\left[\psi'(\mu_i\phi) + \psi'((1-\mu_i)\phi)\right] \cdot \frac{1}{\{g'(\mu_i)\}^2},$$

$c = (c_1, \ldots, c_n)^\top$, com

$$c_i = \phi\left[\psi'\left(\mu_i\phi\right)\mu_i - \psi'((1-\mu_i)\phi)\left(1-\mu_i\right)\right],$$

e $D = \mathrm{diag}(d_1, \ldots, d_n)$, em que

$$d_i = \psi'(\mu_i\phi)\mu_i^2 + \psi'((1-\mu_i)\phi)(1-\mu_i)^2 - \psi'(\phi).$$

A matriz K^{-1} é usada para a obtenção das variâncias assintóticas dos estimadores e pode ser calculada por meio de expressões padrão para a inversa de matrizes particionadas. Sob condições regulares, os estimadores de máxima verossimilhança ($\hat{\beta}$ e $\hat{\phi}$) seguem, aproximadamente, uma distribuição normal multivariada:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \sim \mathcal{N}_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1}\right).$$

Os estimadores de máxima verossimilhança de β e ϕ são obtidos das Equações (2.7) e (2.8) igualadas a zero. Entretanto, como não existe uma forma fechada para expressar explicitamente os estimadores a partir das equações anteriores, é necessário utilizar um algoritmo de otimização, como o de Newton-Raphson, para realizar a maximização numérica da função de log-verossimilhança definida em (2.6).

2.1.2 Pacote Computacional

O pacote `betareg`, inicialmente desenvolvido por Simas e Rocha (2006) para o software estatístico R, permite a modelagem de variáveis contínuas no intervalo (0,1) por meio da distribuição Beta reparametrizada em termos da média μ e da precisão ϕ. A modelagem é feita por máxima verossimilhança, com uso de funções de ligação que conectam os parâmetros a preditores lineares. A função principal, `betareg()`, segue a sintaxe padrão da regressão no R e permite especificar fórmulas tanto para a média quanto para a precisão. O ajuste é realizado por otimização numérica com optim(), e o pacote oferece diversos métodos para análise e inferência do modelo ajustado, como `summary()`, `predict()` e `residuals()`.

Com estrutura modular e flexível, o `betareg` é uma ferramenta acessível e amplamente utilizada para modelagem com a distribuição Beta, especialmente quando se busca representar a média e a dispersão separadamente.

2.2 MODELO SIMPLEX

A distribuição Simplex, inicialmente proposta por Barndorff-Nielsen e Jørgensen (1991), é uma família contínua definida no intervalo entre 0 e 1, especialmente indicada para modelar variáveis restritas a essa faixa. Ao contrário de outras distribuições, ela possui uma estrutura que permite relacionar simultaneamente a média e a variância de forma dependente, o que se mostra particularmente útil em situações de variabilidade heterogênea. Essa característica de modelar a dispersão em função da média é semelhante à da distribuição Beta reparametrizada, na qual a variância também está diretamente associada à média.

2.2.1 Caracterização Teórica do Modelo

De acordo com Silva (2015), diz-se que uma variável aleatória Y pertence à classe dos modelos de dispersão se sua densidade de probabilidade pode ser expressa como

$$f(y;\mu,\sigma^2) = a(y,\sigma^2)\,\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\},\ y\in\Theta, \tag{2.9}$$

em que Θ representa o suporte da distribuição, $\mu\in\Omega$ é interpretado como o parâmetro de localização, e $\sigma^2 > 0$ corresponde ao parâmetro de dispersão da distribuição. A

constante $a(y, \sigma^2)$ atua como um fator de normalização que não depende de μ, garantindo que a função densidade esteja devidamente ajustada. A função $d(y; \mu)$, conhecida como componente *deviance*, é definida no produto cartesiano $\Theta \times \Omega$, satisfazendo as condições $d(y, y) = 0$ sempre que $y = \mu \in \Omega$ e $d(y; \mu) > 0$ para os casos em que $y \neq \mu$, caracterizando assim sua interpretação como medida de discrepância entre a observação y e o valor esperado μ.

A função de variância $\text{Var}(\mu)$, associada aos modelos de dispersão, é dada por

$$\text{Var}(\mu) = 2 \left(\frac{\partial^2 d(y; \mu)}{\partial \mu^2} \bigg|_{y=\mu} \right)^{-1},$$

assumindo que a função $d(y; \mu)$ seja contínua e possua derivadas de segunda ordem em relação a μ, além de satisfazer a condição

$$\frac{\partial^2 d(y; \mu)}{\partial \mu^2} \bigg|_{y=\mu} > 0, \text{ para todo } \mu \in (0, 1).$$

Com base em (2.9), se $Y \sim \text{Simplex}(\mu, \phi)$, então a variável aleatória Y segue distribuição Simplex com média $\mu \in (0, 1)$ e parâmetro de precisão $\phi > 0$, em que $\phi = \frac{1}{\sigma^2}$. Sua FDP é definida como:

$$f(y; \mu, \phi) = \left\{ \frac{\phi}{2\pi\{y(1-y)\}^3} \right\}^{\frac{1}{2}} e^{\left\{ -\frac{\phi}{2} d(y;\mu) \right\}}, \quad 0 \leq y \leq 1, \tag{2.10}$$

em que a chamada unidade *deviance* é

$$d(y; \mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}. \tag{2.11}$$

A média e a variância de Y são dadas, respectivamente, por:

$$\mathbb{E}(Y) = \mu,$$
$$\text{Var(Y)} = \mu(1-\mu) - \sqrt{\frac{\phi}{2}} exp\left\{ \frac{\phi}{2\mu^2(1-\mu)^2} \right\} \Gamma\left\{ \frac{1}{2}; \frac{\phi}{2\mu^2(1-\mu)^2} \right\},$$

em que $\Gamma(t; y) = \int_y^\infty y^{t-1} e^{-y} dy$ é a função gama incompleta.

Para ilustrar o comportamento da distribuição Simplex em diferentes configurações de seus parâmetros, especialmente em relação à simetria e à dispersão, são apresentados os gráficos com diferentes valores da média μ e do parâmetro de dispersão ϕ.

A distribuição Simplex é sensível à variação da média: quando $\mu = 0{,}5$, apresenta simetria em torno do centro; para valores menores, tende a ser assimétrica com cauda à direita, e para valores maiores, com cauda à esquerda. Note que, nesse caso central, as curvas de densidade possuem menor altura em comparação aos cenários extremos. Isso se deve à maior dispersão dos valores ao redor da média, o que suaviza o pico da função. Por

outro lado, quando μ se afasta de 0,5, a densidade se concentra em regiões mais estreitas, elevando sua altura. Além disso, a diminuição de ϕ intensifica essa concentração ao redor da média, ou seja, diminui a variabilidade entre os dados.

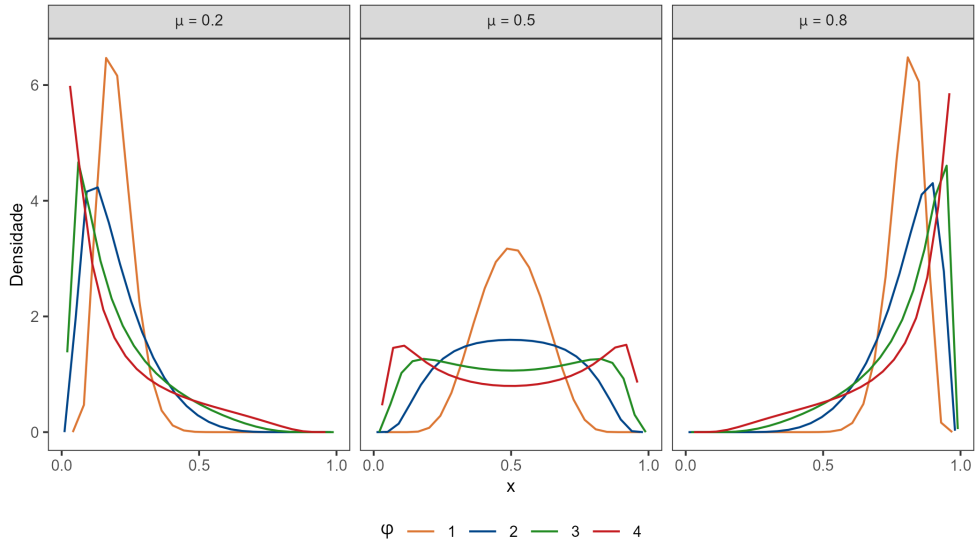

Figura 2 – Densidade da distribuição Simplex para diferentes combinações de μ e ϕ.

Fonte: Elaboração do autor.

Um resultado importante e que reforça a análise gráfica é que, segundo Jørgensen (1997), quando o parâmetro de dispersão tende a zero, entra em cena a teoria assintótica de pequena dispersão, que mostra como essa distribuição passa a se comportar de forma semelhante à distribuição normal. Sendo assim, se $\sigma^2 \to 0$, então

$$\frac{Y-\mu}{\sigma\sqrt{Var(\mu)}} \to N(0,1). \tag{2.12}$$

Para a estimação dos parâmetros realizada via máxima verossimilhança, seguiu-se o mesmo procedimento adotado para a distribuição Beta, modelando-se a média por meio de um preditor linear com função de ligação logito. Seguindo os passos descritos em Fernandes (2019), seja $y_1, \ldots, y_n$ uma a.a. de tamanho N proveniente de uma distribuição Simplex com parâmetros μ e ϕ, cuja função densidade está apresentada em (2.10). A função de verossimilhança correspondente é dada por:

$$\begin{aligned} L(\theta, y) &= \prod_{i=1}^{n} f(y_i \mid \mu, \phi) = \prod_{i=1}^{n} \left\{ \frac{\phi}{2\pi[y_i(1-y_i)]^3} \right\}^{\frac{1}{2}} \exp\left(-\frac{\phi}{2} d(y_i, \mu) \right) \\ &= \left(\frac{\phi}{2\pi \prod_{i=1}^{n}[y_i(1-y_i)]^3} \right)^{\frac{n}{2}} \exp\left(-\frac{\phi}{2} \sum_{i=1}^{n} d(y_i, \mu) \right). \end{aligned}$$

A função log-verossimilhança assume a seguinte forma:

$$l(\theta, y) = log\ L(\theta, y) = \frac{n}{2}\left(log\ \phi - log(2\pi \prod_{i=1}^{n} \{y_i(1-y_i)\}^3) \right) - \frac{\phi}{2} \sum_{i=1}^{n} d(y_i, \mu).$$

Dessa forma, os estimadores de máxima verossimilhança para μ e ϕ são determinados por meio da maximização do logaritmo da função de verossimilhança. Para isso, derivam-se parcialmente $l(\theta, x)$ em relação a μ e a ϕ, respectivamente, obtendo-se os estimadores ao igualar a zero as seguintes equações:

$$\frac{\partial}{\partial \mu} = -\frac{\phi}{2}\sum_{i=1}^{n} d(y_i, \mu),$$
$$\frac{\partial}{\partial \phi} = \frac{n}{2\phi} - \frac{1}{2}\sum_{i=1}^{n} d(y_i, \mu),$$

em que $d(y, \mu)$ é definida em (2.11).

De maneira análoga à Beta, a matriz de informação de Fisher do modelo Simplex pode ser descrita como:

$$K = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix},$$

em que:

$$K_{\beta\beta} = \phi X^\top W X, \quad K_{\beta\phi} = K_{\phi\beta}^\top = X^\top T c, \quad K_{\phi\phi} = \frac{1}{2}\mathbf{1}^\top D \mathbf{1}.$$

Considere

X uma matriz $n \times k$, $T = diag\left\{\frac{1}{g'(\mu_1)}, \ldots, \frac{1}{g'(\mu_n)}\right\}$ e $g'(\mu_i) = \frac{1}{\mu_i(1-\mu_i)}$.

A matriz W é diagonal, $W = \text{diag}(w_1, \ldots, w_n)$, com elementos definidos por:

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{\partial^2}{\partial \mu_i^2}\left(\frac{\phi}{2} d(y_i, \mu_i)\right).$$

O vetor $c = (c_1, \ldots, c_n)^\top$ e a matriz diagonal $D = \text{diag}(d_1, \ldots, d_n)$ possuem os seguintes componentes:

$$c_i = \left(\frac{d}{d\mu_i}\left[\frac{\partial \ell_i}{\partial \phi}\right]\right) \cdot \frac{d\mu_i}{d\eta_i},$$
$$d_i = \mathbb{E}[d(y_i\mu_i)]^2 - [\mathbb{E}(d(y_i, \mu_i))]^2 = Var(d(y_i, \mu_i)).$$

Sob hipóteses regulares, o estimador de máxima verossimilhança $(\hat{\beta}, \hat{\phi})$ é assintoticamente normal com matriz de covariância dada por K^{-1}:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \sim \mathcal{N}_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1}\right).$$

Os estimadores de máxima verossimilhança de β e ϕ não admitem solução analítica explícita, fazendo-se necessário o uso de métodos numéricos para a maximização da função de verossimilhança. Neste trabalho, utilizou-se a função `simplexreg()` do pacote `simplexreg` do R, a qual implementa rotinas de estimação baseadas em máxima verossimilhança por meio de procedimentos numéricos de otimização não linear.

2.2.2 Pacote Computacional

O pacote `simplexreg`, desenvolvido por Zhang et al. (2016), tem como principal objetivo o ajuste de modelos de regressão baseados na distribuição Simplex, sendo especialmente indicado para variáveis contínuas restritas ao intervalo (0,1). Disponível na linguagem R, o pacote oferece ferramentas voltadas à modelagem da média por meio da especificação de funções de ligação apropriadas, com estimação realizada via máxima verossimilhança. Embora também permita a modelagem da dispersão, seu uso é plenamente adequado quando o interesse está centrado na média da variável resposta. Além disso, o pacote disponibiliza recursos para diagnóstico do modelo, bem como funções auxiliares para cálculos relacionados à distribuição simplex, como densidades, quantis e geração de dados aleatórios.

2.3 MODELO NORMAL

A distribuição Normal, também conhecida como Gaussiana, é uma família de distribuições contínuas definida em todo o conjunto dos números reais, amplamente reconhecida por sua relevância teórica e aplicabilidade prática. Caracteriza-se por sua forma simétrica em torno da média e por ser completamente especificada por dois parâmetros: a média (μ) e a variância (σ^2). Sua densidade apresenta uma curva em formato de sino, na qual valores próximos à média são mais prováveis, e a probabilidade decresce suavemente à medida que se afastam do centro. Devido à sua estrutura regular, a Normal é frequentemente empregada como modelo para variáveis contínuas com distribuição aproximadamente simétrica e dispersão constante. Além disso, a distribuição normal com média zero e desvio padrão igual a um é conhecida como distribuição normal padrão.

A distribuição de Gauss ocupa um papel central na teoria das probabilidades e na estatística devido ao Teorema Central do Limite, que estabelece que a média de um grande número N de variáveis aleatórias independentes e identicamente distribuídas tende a seguir uma distribuição normal, independentemente da forma da distribuição original dessas variáveis, como mostrado em (2.12). Isso ocorre porque, à medida que o número de observações aumenta, os efeitos individuais das variáveis se combinam de maneira que o comportamento agregado se aproxima de uma curva simétrica em formato de sino, característica da distribuição normal. Desse modo, a distribuição normal serve como ferramenta para modelar o resultado acumulado de diferentes processos estocásticos.

2.3.1 Caracterização Teórica do Modelo

A FDP da distribuição Normal, para uma variável aleatória contínua $Y \sim N(\mu, \sigma^2)$, é dada pela seguinte expressão:

$$f(y; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \; y \in \mathbb{R},$$

em que $\mu \in \mathbb{R}$ é o parâmetro de localização e $\sigma^2 > 0$ é o parâmetro de escala.

Considerando o parâmetro ϕ para similaridade aos demais modelos, seu valor esperado e variância são dados por:

$$\mathbb{E}(Y) = \mu,$$
$$\text{Var}(Y) = \sigma^2 = \phi^2.$$

Dessa forma, considerando a média restrita ao intervalo unitário e diferentes valores para a variância, é apresentado o gráfico com o comportamento da distribuição. Note que os valores que se estendem nas extremidades do eixo x apresentam densidade praticamente nula em todos os gráficos, o que indica que essas regiões têm baixa probabilidade de ocorrência. Isso mostra que a distribuição está fortemente concentrada ao redor da média, especialmente quando o parâmetro ϕ é pequeno. À medida que ϕ aumenta, a densidade se espalha um pouco mais, suavizando a curva e tornando as caudas ligeiramente mais espessas, mas ainda assim com baixa densidade nas extremidades.

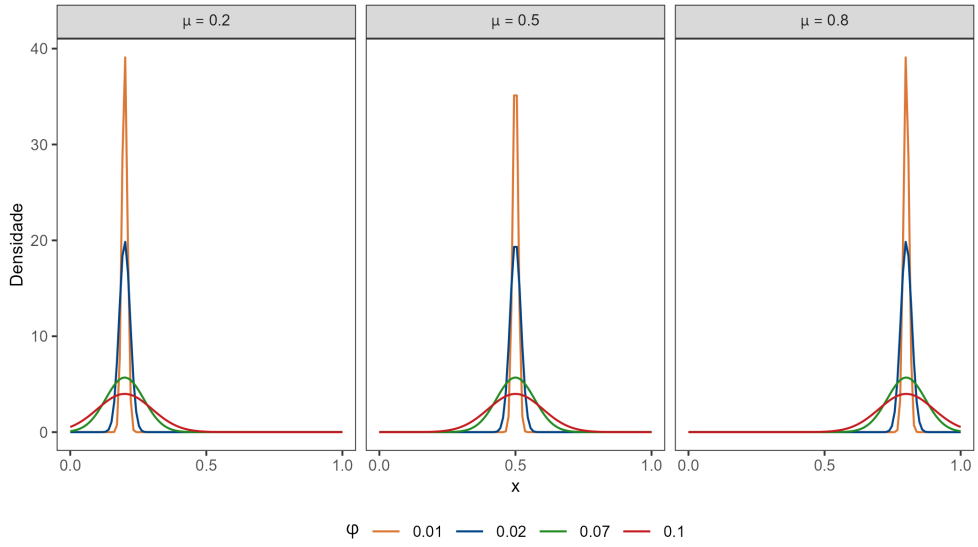

Figura 3 – Densidade da distribuição Normal para diferentes combinações de μ e ϕ.
Fonte: Elaboração do autor.

A estimação dos parâmetros foi realizada por meio do método da máxima verossimilhança. De forma análoga às distribuições Beta e Simplex, a média foi modelada por meio de um preditor linear. Mas, diferentemente dessas distribuições, adotou-se a função de ligação identidade. Sendo assim, a média é diretamente igual ao preditor linear, ou seja:

$$g(\mu_i) = \mu_i = \eta_i = x_i^\top \beta.$$

Neste caso, a função identidade atua como a função de ligação canônica para a distribuição Normal, o que simplifica a modelagem. Isso porque a relação entre a média

da variável resposta e os preditores é direta e linear, sem transformações intermediárias. Como consequência, a interpretação dos coeficientes do modelo torna-se mais intuitiva: cada parâmetro representa a variação esperada na média da resposta para uma unidade de alteração na respectiva variável preditora, mantendo as demais constantes.

Por McCullagh e Nelder (1989), considere uma amostra aleatória $y_1, y_2, \ldots, y_n$ proveniente de uma distribuição Normal com média μ e variância σ^2. A função de verossimilhança é dada por:

$$L(\mu, \sigma^2; y) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2} \right) \right].$$

Tomando o logaritmo da função de verossimilhança, obtém-se:

$$l(\mu, \sigma^2; y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.$$

A diferenciação do logaritmo da função de verossimilhança em relação aos parâmetros μ e σ^2 resulta nas derivadas, que, para encontrar as estimativas que maximizam a função, são igualadas a zero:

$$\frac{\partial}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \;\Rightarrow\; \hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$
$$\frac{\partial}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \mu)^2 \;\Rightarrow\; \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2.$$

Essas equações correspondem às condições de máxima verossimilhança para os parâmetros da distribuição Normal, resultando nos estimadores clássicos da média e da variância populacional não corrigida. A especificação do modelo está, assim, completa ao se definir essa estrutura sistemática juntamente com a suposição de normalidade dos erros com variância constante σ^2 ou seja:

$$Y \sim N(X\beta, \sigma^2 I),$$

em que β é o vetor de parâmetros do modelo e I é a matriz identidade de dimensão $n \times n$, sendo n o número de observações.

A matriz de informação de Fisher, que corresponde à esperança da segunda derivada negativa da função de log-verossimilhança, desempenha papel fundamental tanto na caracterização das propriedades assintóticas dos estimadores quanto na implementação de métodos iterativos de estimação. No contexto do modelo normal com ligação identidade, essa matriz pode ser escrita como:

$$K = \begin{pmatrix} \sigma^{-2} X^\top X & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}.$$

Note que os termos cruzados são nulos, refletindo a independência (no sentido de informação) entre os estimadores de β e σ^2 sob o modelo normal. Essa estrutura diagonal facilita o cálculo da inversa da matriz de informação, que é usada para derivar a matriz de covariância assintótica dos estimadores de máxima verossimilhança:

$$K^{-1} = \begin{pmatrix} \sigma^2(X^\top X)^{-1} & 0 \\ 0 & (2\sigma^4)/n \end{pmatrix}. \tag{2.13}$$

Como os termos cruzados da matriz são nulos, as estimativas de máxima verossimilhança β e σ^2 são assintoticamente independentes, ou seja, a informação que uma traz não interfere na outra na amostra grande.

Por (2.13), conclui-se que

$$\hat{\beta} \sim N_p\Big(\beta,\ \sigma^2(X^\top X)^{-1}\Big)$$
$$\hat{\sigma^2} \sim N\left(\sigma^2,\ \frac{2\sigma^4}{n}\right)$$

A diagonalização da matriz de informação garante que $\hat{\beta}$ e $\hat{\sigma}^2$ sejam independentes para todo n, um resultado mais forte que a simples independência assintótica.

2.3.2 Pacote Computacional

A função `glm()` pertence ao pacote stats, que é um dos pacotes base do R e já vem instalado automaticamente com a linguagem. Trata-se de uma ferramenta nativa amplamente utilizada para ajustar modelos lineares generalizados (Generalized Linear Models). Através da especificação da família de distribuições e da função link adequada, o `glm()` permite modelar diferentes tipos de respostas, como Binomial, Poisson, Gamma, entre outras, além da distribuição Normal. Essa versatilidade torna o `glm()` extremamente útil para análises estatísticas que envolvem dados categóricos, contagens ou variáveis contínuas com diversas distribuições. Utilizando o método de máxima verossimilhança, a função estima os coeficientes do modelo, oferecendo base para inferência, previsão e interpretação dos efeitos das variáveis explicativas.

3 SIMULAÇÃO

Neste capítulo, são realizados estudos de simulação utilizando experimentos com diferentes distribuições aplicadas à modelagem de uma variável contínua limitada. Amostras são geradas em diferentes tamanhos e as estimativas dos parâmetros dos modelos são avaliadas por meio de medidas de desempenho. Os resultados são apresentados em tabelas e gráficos, permitindo a comparação do desempenho das distribuições consideradas.

De acordo com Donatelli e Konrath (2005), a Simulação de Monte Carlo (SMC) é um método que consiste em simular repetidas amostras aleatórias segundo algum tipo de distribuição pré-especificada, para que, em seguida, seja possível analisar conjuntamente o resultado obtido a partir de cada uma dessas amostras. Assim, em vez de apenas coletar dados, criar uma hipótese e testá-la, as simulações nos possibilitam gerar os valores, definindo previamente como eles devem se comportar. Dessa maneira, os números são obtidos utilizando sorteios no software estatístico R. Por Júnior et al. (2011), a cada iteração, o resultado é armazenado e, ao final de todas as repetições, a sequência de resultados gerados é transformada em uma distribuição de frequência que possibilita calcular estatísticas descritivas, como média, valor mínimo, valor máximo e desvio-padrão. Com base no livro de Mooney (1997), uma das principais vantagens do método é a possibilidade de simular um experimento nos moldes tradicionais, já que permite controlar tanto o número de simulações quanto as características das variáveis envolvidas. Ao repetir a geração dos dados com o uso da aleatoriedade, é possível analisar como pequenas variações no processo impactam - ou não - os resultados, permitindo avaliar a sensibilidade da simulação a mudanças sutis em sua implementação. O procedimento básico de Monte Carlo é seguir os seguintes passos:

1. Especificar a pseudopopulação;
2. Gerar amostras da pseudopopulação;
3. Calcular o estimador na pseudoamostra;
4. Repetir várias vezes os passos 2 e 3;
5. Construir a distribuição amostral do estimador.

O prefixo 'pseudo' destaca que esses elementos são aproximações artificiais, criadas para estudar propriedades estatísticas, e não dados diretamente observados no mundo real. Ampliando a explicação, a pseudopopulação Beta é definida simbolicamente pelo modelo:

$$Y_i \sim \text{Beta}(\mu_i\phi, (1-\mu_i\phi)), \quad \mu_i = \frac{\exp\{\eta_i\}}{1-\exp\{\eta_i\}}, \quad \eta_i = x_i^\top \beta, \tag{3.1}$$

em que $\beta = [1, 1]^\top$ é o vetor de coeficientes verdadeiros, $\phi = 10$ é o parâmetro de dispersão e x_i é a matriz de covariáveis com a primeira coluna de 1 (intercepto) e uma variável uniforme U(0,1). O modelo é implementado computacionalmente usando a função `betareg()` do pacote `betareg` e a função `rbeta()` para gerar os dados simulados.

A pseudopopulação Simplex é definida por:

$$Y_i \sim \text{Simplex}(\mu_i, \phi), \quad \mu_i = \frac{\exp\{\eta_i\}}{1 - \exp\{\eta_i\}}, \quad \eta_i = x_i^\top \beta, \tag{3.2}$$

em que $\beta = [1, 1]^\top$, $\phi = 1$ e x_i é gerado com distribuição uniforme em (0,1). O modelo é implementado via software R usando a função `simplexreg()` do pacote `simplexreg` e a função `rsimplex()` para geração de dados simulados.
A última pseudopopulação é definida pelo modelo de regressão linear com erros normais:

$$Y_i \sim \text{N}(\mu_i, \phi^2), \quad \mu_i = \eta_i = x_i^\top \beta, \tag{3.3}$$

em que $\beta = [1, 1]^\top$, $\phi = 0{,}1$ é o desvio padrão fixo da distribuição normal (ou seja, variância $\phi^2 = 0{,}01$), x_i é gerado com distribuição U(0,1). A variável resposta é simulada com `rnorm()` e o modelo ajustado com `glm()`, ambos do R base.

Sob esta abordagem, realiza-se uma SMC com 5.000 réplicas para avaliar o desempenho dos estimadores nos três modelos de regressão. Foram considerados diferentes tamanhos de amostra, especificamente 10, 20, 40, 80 e 160 observações, com o objetivo de verificar como a precisão dos estimadores evolui conforme a quantidade de dados aumenta.

Para cada tamanho amostral, gera-se um conjunto de covariáveis simuladas, incluindo o intercepto, que são usadas para calcular o valor esperado da variável resposta μ_i. Como apresentado em (3.1) e (3.2), nos modelos Beta e Simplex, a relação entre μ_i e as covariáveis é feita por meio da função logística, conforme o modelo:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_i,$$

em que β_0 e β_1 são os coeficientes do modelo e x_i representa a covariável associada à observação i.

No modelo Normal, conforme (3.3), a função de ligação utilizada é a identidade, ou seja, a média da variável resposta é modelada diretamente como

$$\mu_i = \beta_0 + \beta_1 x_i.$$

São utilizadas as funções de ligação logito e identidade devido à sua interpretação direta e intuitiva dos coeficientes. A função logito transforma a média da resposta (restrita entre 0 e 1) em uma escala ilimitada. Nessa escala, os coeficientes indicam o efeito de uma covariável sobre o log-odds da média. Assim, aumentos nas covariáveis implicam aumentos (ou reduções) proporcionais na razão entre a média e seu complemento, facilitando a interpretação em termos de tendências. A função identidade, usada com a distribuição Normal, mantém a média da resposta diretamente igual à combinação linear das covariáveis. Com isso, cada coeficiente representa o quanto a média da resposta muda quando a covariável correspondente aumenta uma unidade, mantendo as demais constantes.

Com estas médias e parâmetros de precisão fixos, específicos para cada modelo, são gerados dados sintéticos respeitando as respectivas distribuições e parametrizações. Em cada replicação, o modelo é ajustado aos dados simulados e as estimativas dos parâmetros são obtidas. Ao final de todas as replicações, calcula-se, para cada parâmetro, a média das estimativas, que representa o valor central obtido ao longo das simulações. O erro padrão é obtido a partir do desvio padrão dessas estimativas, refletindo a variabilidade esperada sob repetição do experimento. O viés é calculado como a diferença entre a média das estimativas e o valor verdadeiro do parâmetro, sendo uma medida de tendência sistemática de erro. O viés absoluto considera apenas a magnitude dessa diferença, sem levar em conta sua direção, e fornece uma medida geral do desvio médio. Por fim, o erro quadrático médio (EQM) é obtido pela média dos quadrados das diferenças entre as estimativas e o valor verdadeiro, combinando em uma única métrica tanto a variabilidade quanto o viés das estimativas.

Os resultados dessas métricas são organizados em tabelas que permitem uma análise comparativa do desempenho dos estimadores nos diferentes modelos e condições amostrais. Esse procedimento possibilita observar o comportamento dos estimadores frente à variabilidade inerente ao processo amostral.

3.1 ESTUDO DE SIMULAÇÃO DO MODELO BETA

N	Estimador	Média	EP	Viés	|Viés|	EQM
10	$\hat{\beta}_1$	1.022	0.387	0.022	0.308	0.150
	$\hat{\beta}_2$	1.025	1.302	0.025	1.034	1.696
	$\hat{\phi}$	16.874	11.669	6.874	7.973	183.399
20	$\hat{\beta}_1$	1.007	0.323	0.007	0.257	0.104
	$\hat{\beta}_2$	1.029	0.617	0.029	0.489	0.381
	$\hat{\phi}$	12.476	4.567	2.476	3.612	26.983
40	$\hat{\beta}_1$	1.012	0.285	0.012	0.226	0.082
	$\hat{\beta}_2$	1.000	0.470	0.000	0.374	0.221
	$\hat{\phi}$	11.133	2.722	1.133	2.163	8.693
80	$\hat{\beta}_1$	0.999	0.155	-0.001	0.123	0.024
	$\hat{\beta}_2$	1.007	0.269	0.007	0.216	0.072
	$\hat{\phi}$	10.526	1.741	0.526	1.387	3.307
160	$\hat{\beta}_1$	1.003	0.106	0.003	0.085	0.011
	$\hat{\beta}_2$	0.999	0.200	-0.001	0.160	0.040
	$\hat{\phi}$	10.258	1.162	0.258	0.936	1.417

Tabela 1 – Desempenho de parâmetros do modelo Beta em simulações Monte Carlo.

3.2 ESTUDO DE SIMULAÇÃO DO MODELO SIMPLEX

N	Estimador	Média	EP	Viés	|Viés|	EQM
10	$\hat{\beta}_1$	1.006	0.222	0.006	0.178	0.049
	$\hat{\beta}_2$	1.017	0.675	0.017	0.538	0.456
	$\hat{\phi}$	0.993	0.493	-0.007	0.387	0.243
20	$\hat{\beta}_1$	1.007	0.195	0.007	0.156	0.038
	$\hat{\beta}_2$	0.992	0.281	-0.008	0.225	0.079
	$\hat{\phi}$	0.997	0.339	-0.003	0.267	0.115
40	$\hat{\beta}_1$	1.003	0.125	0.003	0.099	0.016
	$\hat{\beta}_2$	1.001	0.193	0.001	0.154	0.037
	$\hat{\phi}$	1.006	0.229	0.006	0.182	0.052
80	$\hat{\beta}_1$	1.004	0.087	0.004	0.070	0.008
	$\hat{\beta}_2$	0.996	0.140	-0.004	0.112	0.020
	$\hat{\phi}$	1.002	0.162	0.002	0.129	0.026
160	$\hat{\beta}_1$	1.001	0.064	0.001	0.051	0.004
	$\hat{\beta}_2$	1.000	0.102	0.000	0.081	0.010
	$\hat{\phi}$	1.000	0.116	0.000	0.093	0.013

Tabela 2 – Desempenho de parâmetros do modelo Simplex em simulações Monte Carlo.

3.3 ESTUDO DE SIMULAÇÃO DO MODELO NORMAL

N	Estimador	Média	EP	Viés	|Viés|	EQM
10	$\hat{\beta}_1$	1.000	0.055	0.000	0.044	0.003
	$\hat{\beta}_2$	1.003	0.178	0.003	0.143	0.032
	$\hat{\phi}$	0.097	0.024	-0.003	0.020	0.001
20	$\hat{\beta}_1$	1.000	0.036	0.000	0.029	0.001
	$\hat{\beta}_2$	1.001	0.076	0.001	0.061	0.006
	$\hat{\phi}$	0.098	0.016	-0.002	0.013	0.000
40	$\hat{\beta}_1$	1.000	0.038	0.000	0.030	0.001
	$\hat{\beta}_2$	0.999	0.055	-0.001	0.044	0.003
	$\hat{\phi}$	0.100	0.011	0.000	0.009	0.000
80	$\hat{\beta}_1$	1.000	0.023	0.000	0.018	0.001
	$\hat{\beta}_2$	1.000	0.042	0.000	0.034	0.002
	$\hat{\phi}$	0.100	0.008	0.000	0.006	0.000
160	$\hat{\beta}_1$	1.000	0.016	0.000	0.013	0.000
	$\hat{\beta}_2$	1.001	0.027	0.001	0.021	0.001
	$\hat{\phi}$	0.100	0.006	0.000	0.005	0.000

Tabela 3 – Desempenho de parâmetros do modelo Normal em simulações Monte Carlo.

3.4 RESULTADOS COMPARATIVOS DAS SIMULAÇÕES

Os resultados das Tabelas 1, 2 e 3 mostram que o modelo Normal apresenta o melhor desempenho entre os três analisados, com viés praticamente nulo e erros padrão muito baixos para todos os parâmetros. O modelo Simplex aparece em segundo lugar, exibindo um viés pequeno e erros padrão um pouco maiores, além de erros quadráticos médios inferiores aos do modelo Beta, mas ainda superiores aos do Normal. Já o modelo Beta tem o desempenho mais fraco, com viés elevado, especialmente no parâmetro ϕ, além de erros padrão e erros quadráticos médios significativamente maiores, demonstrando maior variabilidade e menor precisão nas estimativas. À medida que o tamanho da amostra aumenta, todos os modelos melhoram seus resultados, porém o Normal estabiliza seus erros rapidamente em níveis baixos, o Simplex melhora de forma significativa, mas não alcança a precisão do Normal, e o Beta reduz seus erros mais lentamente, mantendo valores relativamente altos mesmo com amostras maiores.

Após a apresentação das tabelas com as métricas de desempenho dos parâmetros estimados nos modelos simulados, serão exibidos gráficos comparativos dessas métricas, erro padrão, viés absoluto e erro quadrático médio, para as três distribuições, contemplando os diferentes tamanhos amostrais. Esses gráficos visam reforçar a análise e evidenciar as tendências já observadas nas tabelas.

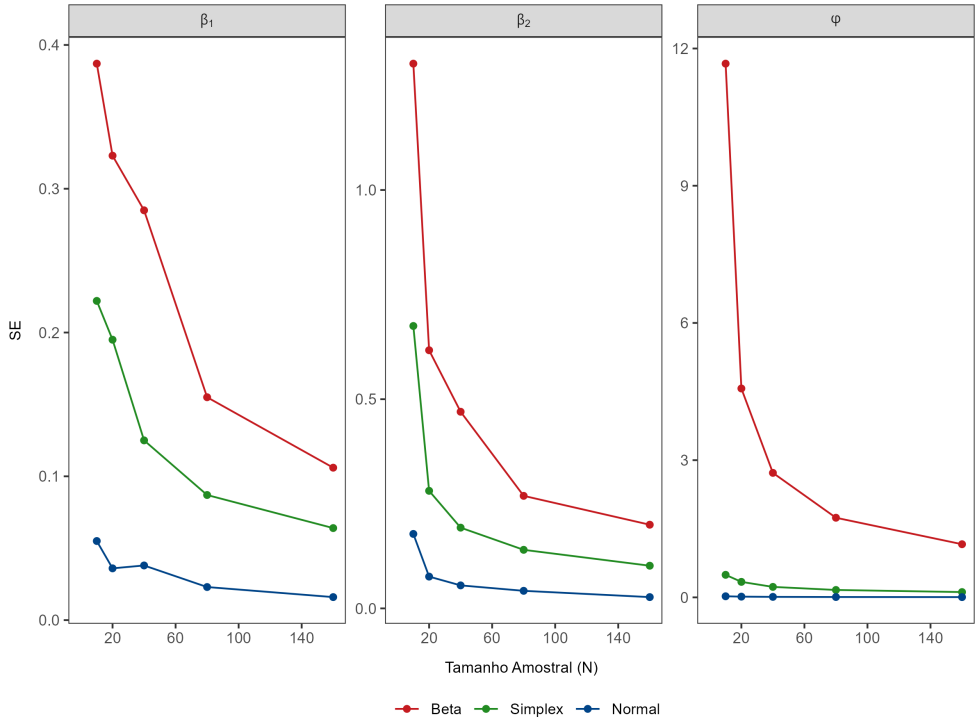

Figura 4 – Análise do erro padrão sob variação de modelos e amostras.

Fonte: Elaboração do autor.

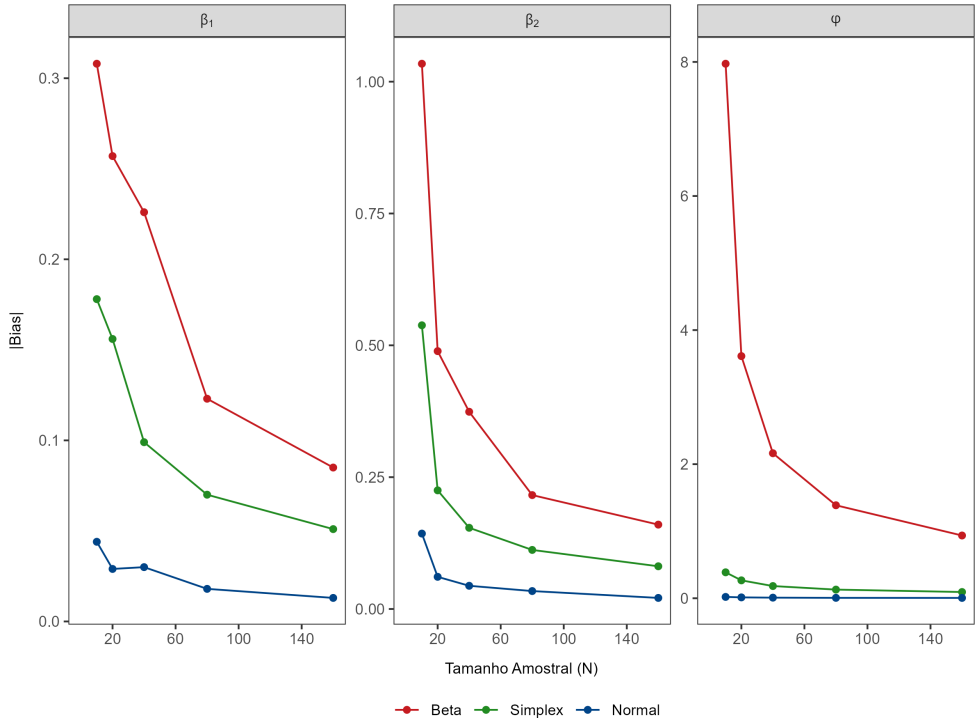

Figura 5 – Análise do viés absoluto sob variação de modelos e amostras.

Fonte: Elaboração do autor.

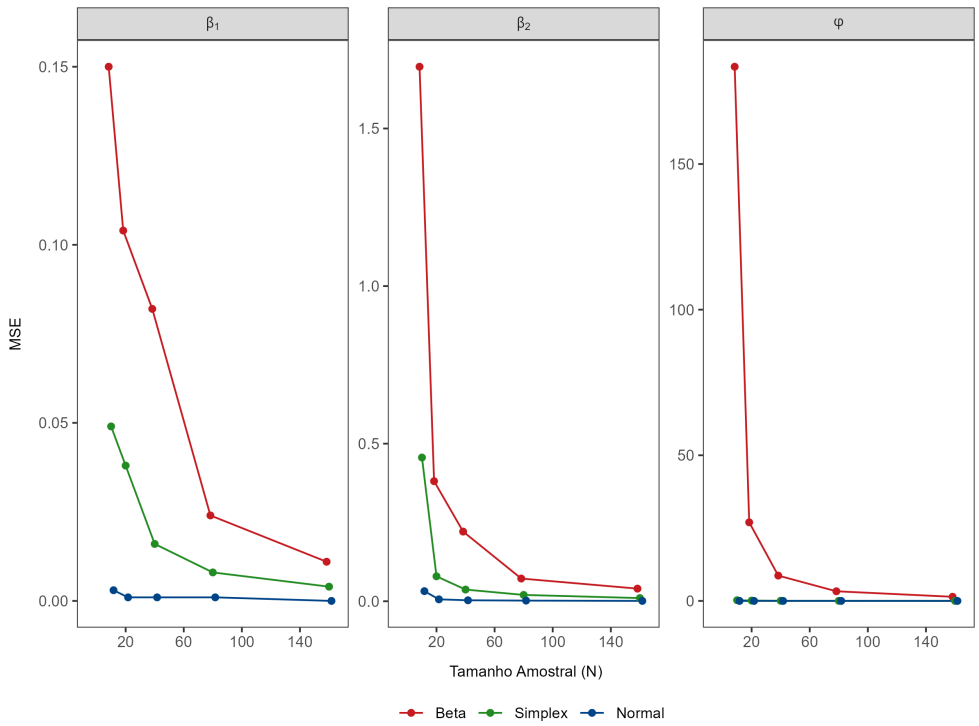

Figura 6 – Análise do erro quadrático médio sob variação de modelos e amostras.

Fonte: Elaboração do autor.

4 APLICAÇÃO

Para verificar a efetividade da metodologia desenvolvida, optou-se por aplicá-la a uma base de dados reais. Neste capítulo, é explorada uma aplicação, com o objetivo de examinar os resultados obtidos na prática.

4.1 CAMPEÕES BRASILEIROS

Entre os anos de 2003 e 2024, foi reunido um conjunto de informações sobre os times campeões da primeira divisão do futebol brasileiro masculino, contemplando algumas características dessas equipes vencedoras. A competição, nomeada como Brasileirão, foi reconhecida como a liga nacional mais forte do mundo em 2021 e 2022 pela International Federation of Football History and Statistics (IFFHS). Como todo ano há um novo campeão, a base de dados é continuamente atualizada com novas informações, refletindo os perfis das equipes mais recentes. Por isso, um modelo previamente definido pode não se manter ideal nos anos seguintes, uma vez que novos campeões podem apresentar características diferentes das equipes anteriores ou até mesmo dos seus próprios clubes, já que jogadores podem ser vendidos ou contratados de uma temporada para outra.

Apresentada em Magalhães et al. (2025), esta base de dados tem 22 observações com 26 variáveis cada sobre o campeão brasileiro de cada um dos anos. As variáveis são: o ano da competição, time campeão, abreviação do nome do clube, cidade do time campeão, estado do time campeão, se o time foi ou não campeão estadual, número de clubes brasileiros no final da CONMEBOL Libertadores, um time brasileiro foi ou não campeão da CONMEBOL Libertadores, número de clubes brasileiros nas finais da CONMEBOL Libertadores e CONMEBOL Sudamericana, se o clube possui estádio próprio, se é uma temporada de Copa do Mundo da FIFA, quantos clubes jogam no mesmo estádio, número de rivais, se era ou não o atual campeão, número de participantes no campeonato, número de gols na competição, média de gols por partida na competição, pontos obtidos, jogos disputados, número de vitórias, número de empates, número de derrotas, número de gols marcados, número de gols sofridos, saldo de gols e aproveitamento (razão entre os pontos obtidos e o triplo de jogos disputados). No formato adotado pela competição, cada equipe enfrenta todas as demais em jogos de ida e volta. A pontuação segue a seguinte lógica: três pontos são atribuídos por vitória, um ponto por empate e nenhum ponto em caso de derrota. Ao término do campeonato, o título é concedido ao clube que acumular a maior pontuação.

O objetivo desta aplicação é investigar se o aproveitamento dos times campeões pode ser explicado por alguma(s) das variáveis consideradas. Para isso, serão inicialmente apresentadas algumas informações descritivas sobre os dados, com uma análise exploratória que possibilite uma melhor compreensão dos padrões observados. A Figura 7 apresenta o

histograma do aproveitamento dos campeões e evidencia uma distribuição assimétrica. A média de aproveitamento foi de 0,67, destacada pela linha vertical. O gráfico tem limites do eixo x entre 0 e 1, de modo a ressaltar que o aproveitamento está contido em um intervalo unitário. Nota-se que os valores se agrupam principalmente entre 0,65 e 0,75, aproximadamente, sem a ocorrência de outliers.

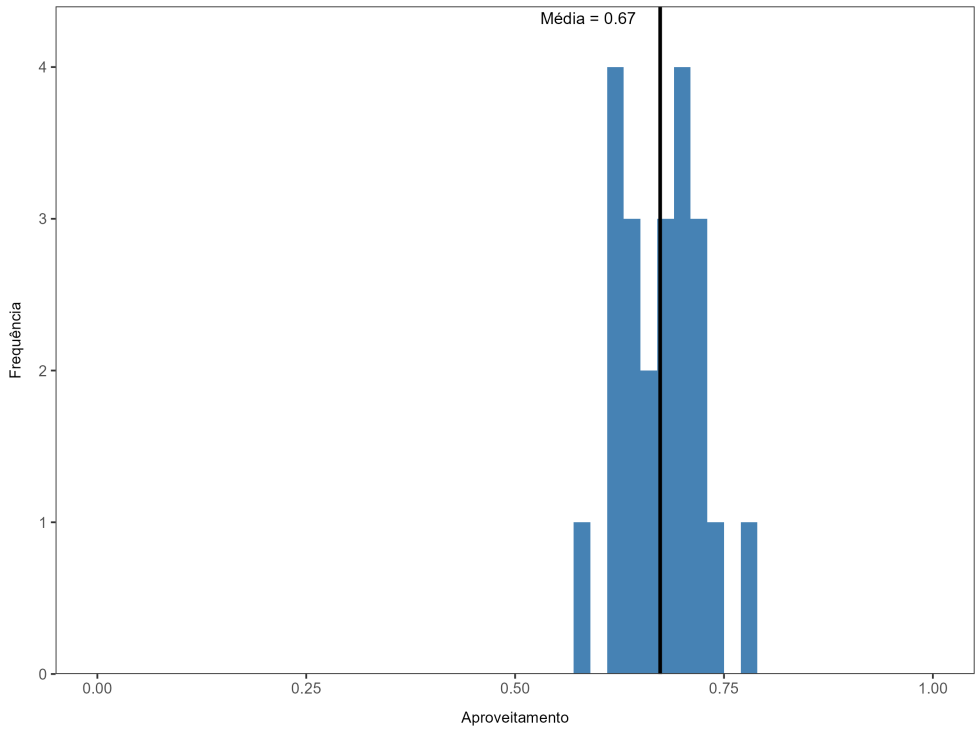

Figura 7 – Histograma do aproveitamento.

Fonte: Elaboração do autor.

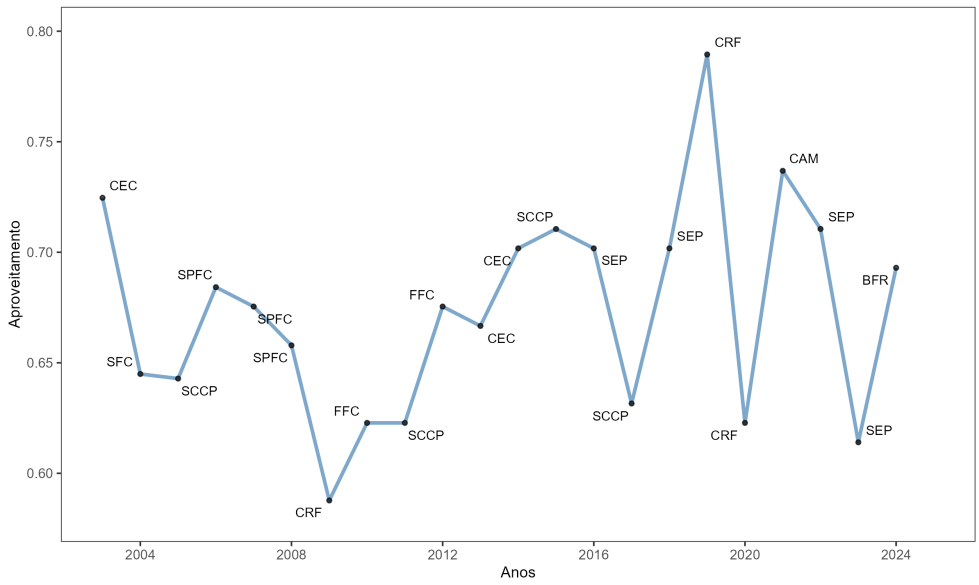

Figura 8 – Aproveitamento de pontos dos times campeões ao longo dos anos.

Fonte: Elaboração do autor.

A Figura 8 mostra o aproveitamento de pontos dos campeões do Campeonato Brasileiro ao longo dos anos, possibilitando comparações consistentes graças ao sistema de pontos corridos. Os índices variaram entre aproximadamente 0,58 e 0,80, com o maior aproveitamento registrado em 2019 pelo Flamengo. Anos como 2009 e 2023 apresentaram os menores índices, abaixo de 0,60. Clubes como Palmeiras, Flamengo, Corinthians e Cruzeiro se destacam pela frequência dos títulos, evidenciando certa hegemonia, embora sem uma tendência clara de melhora ou queda no desempenho, refletindo uma competitividade instável no torneio.

Outro aspecto a ser explorado é o número de títulos conquistados por Estado, com São Paulo liderando com 12 títulos, seguido pelo Rio de Janeiro, que acumulou 6 conquistas, e Minas Gerais, com 4 títulos; esses dados evidenciam a predominância histórica dos estados do Sudeste no cenário nacional, refletindo a força e a tradição das equipes dessas regiões ao longo do tempo. A Figura 9 resume essas análises.

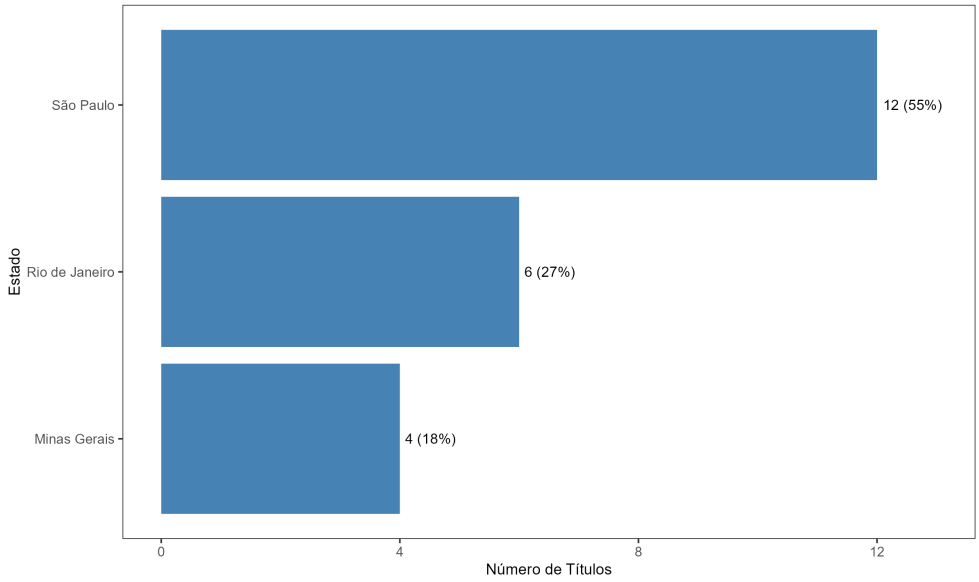

Figura 9 – Número de títulos por Estado.

Fonte: Elaboração do autor.

A Figura 10 exibe o boxplot do aproveitamento de pontos dos times campeões por Estado. A análise indica que o Rio de Janeiro apresenta maior variabilidade, incluindo um outlier, com mediana próxima a 0,65. Em contraste, Minas Gerais mostra menor variabilidade e uma mediana acima de 0,7, embora ambos os estados tenham poucos registros, o que torna seus resultados mais suscetíveis a variações. Já São Paulo, com um número maior de observações, apresenta baixa variabilidade e mediana ligeiramente superior a 0,65.

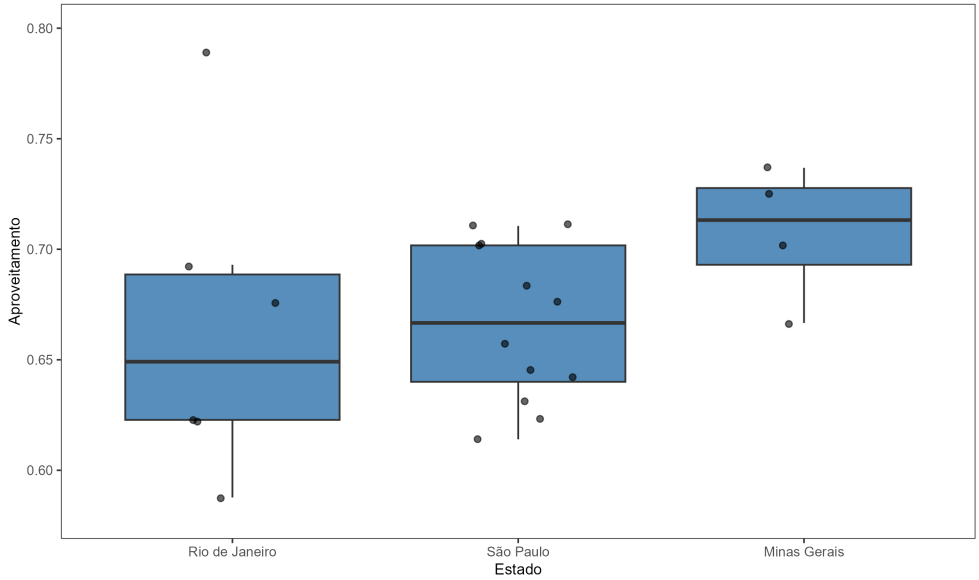

Figura 10 – Aproveitamento de pontos dos times campeões por Estado.

Fonte: Elaboração do autor.

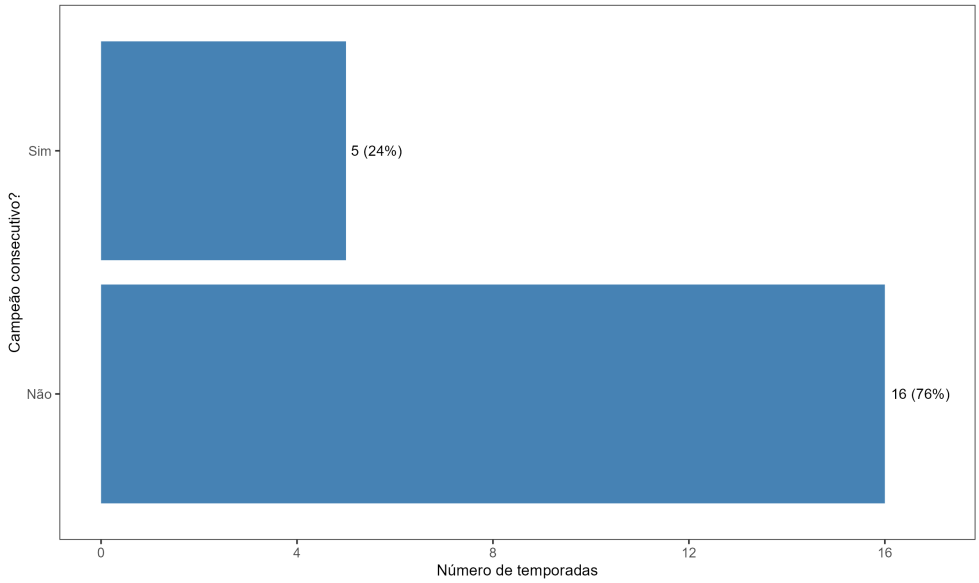

Figura 11 – Número de clubes com títulos brasileiros consecutivos.

Fonte: Elaboração do autor.

A Figura 11 apresenta o número de clubes vencedores por 2 anos consecutivos. A análise mostra que é incomum o campeão repetir o título na temporada seguinte.

A seguir, a matriz de correlações das variáveis numéricas da base de dados apresenta os coeficientes de correlação de Pearson entre cada par de variáveis, permitindo identificar a intensidade e a direção das relações lineares entre elas.

A análise da matriz de correlação mostra que a variável Ratio (aproveitamento) apresenta relação fraca ou inexistente com a maioria das demais variáveis. Com Contfin (número de clubes brasileiros nas finais de torneios continentais) e Num (número de clubes no torneio), a correlação é praticamente nula, indicando ausência de associação linear. A relação com Libfinal (número de clubes brasileiros na final da CONMEBOL), Rivals (número de clubes considerados rivais na competição), Staterivals (número de clubes do mesmo estado), Ga (gols contra) e Goals (total de gols) é negativa e fraca, sugerindo que essas variáveis possuem influência limitada sobre o Ratio. Já a variável Gav (média de gols por jogo) apresenta uma correlação negativa moderada com o Ratio, indicando que valores mais altos dessa média tendem a estar associados a menores aproveitamentos. Em contraste, a variável Gf (gols a favor) tem uma correlação positiva fraca, sugerindo que um maior número de gols a favor pode estar levemente relacionado a um maior aproveitamento. A única correlação forte observada é com Gd (saldo de gols, ou seja, gols a favor menos gols contra), com valor de 0,72, indicando que equipes com maior saldo de gols tendem a apresentar maior aproveitamento.

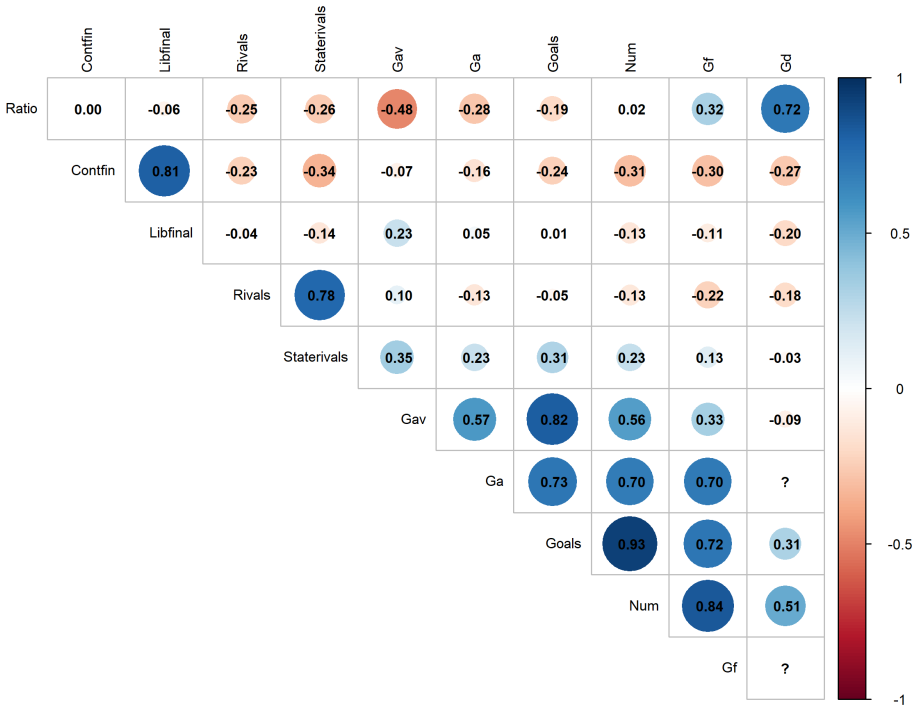

Figura 12 – Matriz de correlação.

Fonte: Elaboração do autor.

A Figura 13 reforça a relação direta entre o saldo de gols e o aproveitamento dos times campeões, mostrando que um aumento no saldo corresponde a um desempenho melhor.

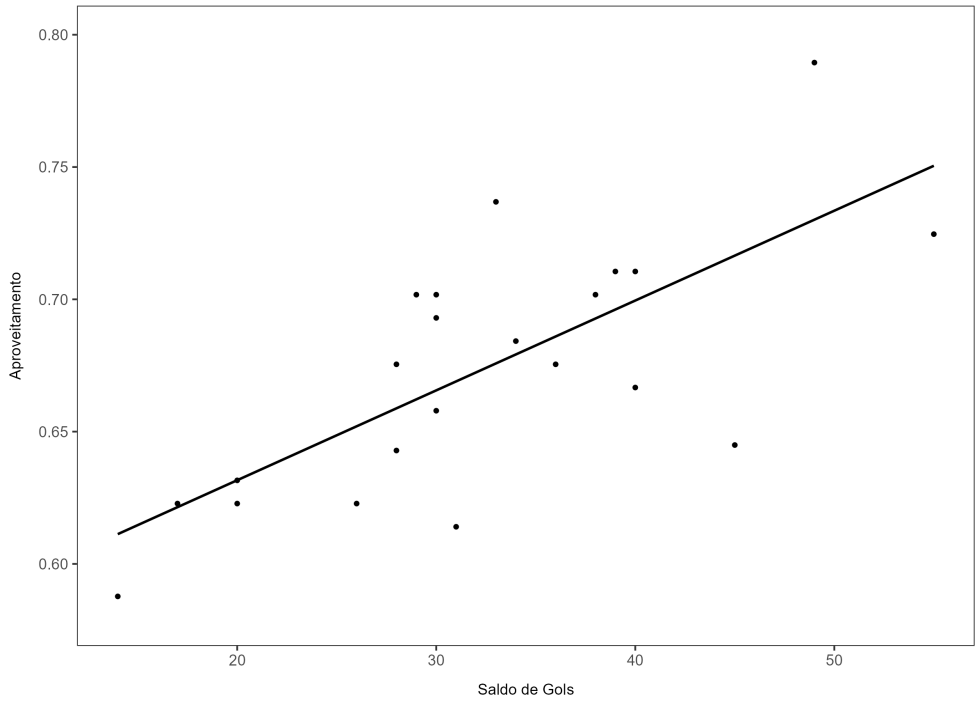

Figura 13 – Relação entre o Saldo de Gols e o Aproveitamento.

Fonte: Elaboração do autor.

Antes do ajuste dos modelos estatísticos, é realizada a análise de autocorrelação e autocorrelação parcial do aproveitamento de pontos dos campeões, com o objetivo de verificar a existência de dependência temporal entre as observações. Como os modelos estudados pressupõem independência entre os dados, essa etapa valida a adequação desses modelos ao contexto.

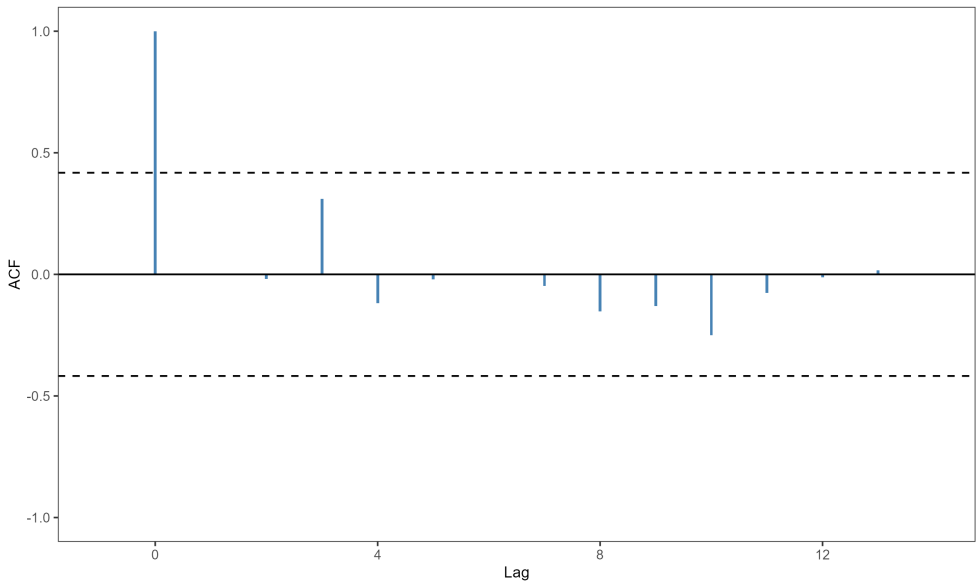

Figura 14 – Autocorrelação do aproveitamento.

Fonte: Elaboração do autor.

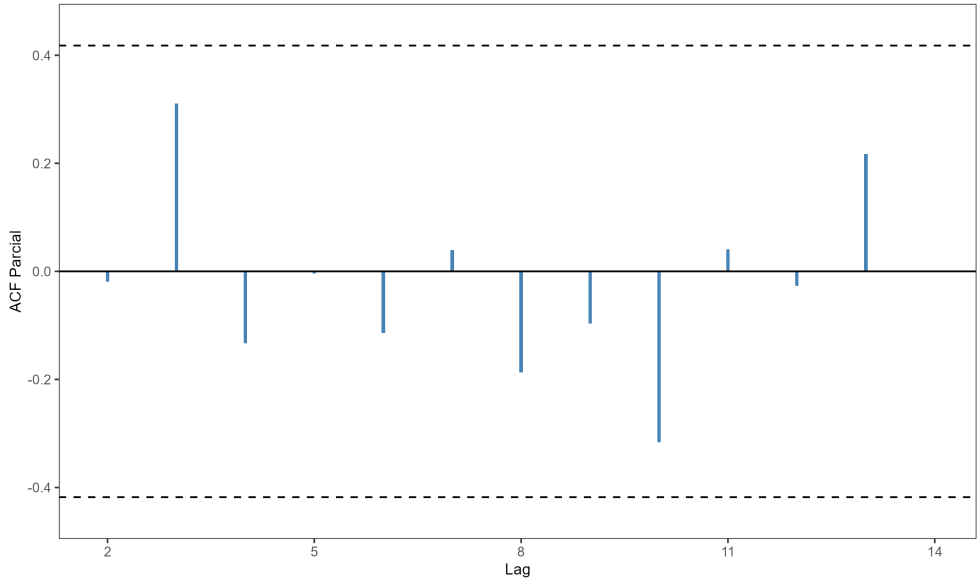

Figura 15 – Autocorrelação parcial do aproveitamento.

Fonte: Elaboração do autor.

Os resultados indicaram ausência de autocorrelação significativa, o que justifica a aplicação de modelos que assumem independência. Isso ocorre porque os coeficientes das Figuras 14 e 15 se mantiveram dentro dos limites de confiança, sugerindo que os valores da série não estão correlacionados entre si em diferentes defasagens.

Diante disso, para analisar a relação entre o aproveitamento e características da trajetória do clube vencedor, ajustam-se modelos de regressão baseados nas distribuições Beta, Simplex e Normal pelo software estatístico R, conforme a seguinte forma funcional:

$$\eta^\top = \beta_0^\top + \beta_1^\top Gd,$$

em que o vetor de preditores é $\eta^\top = \begin{cases} \eta_B = \text{aproveitamento com o modelo Beta,} \\ \eta_S = \text{aproveitamento com o modelo Simplex,} \\ \eta_N = \text{aproveitamento com o modelo Normal,} \end{cases}$

$\beta_0^\top$ é o vetor de coeficientes do intercepto e
$\beta_1^\top$ é o vetor de coeficientes de regressão relacionados ao saldo de gols.

A escolha do saldo de gols como única variável explicativa fundamenta-se na sua correlação significativa com a variável resposta, além do fato de que as demais variáveis analisadas não apresentaram evidências estatísticas de relevância no modelo. Dessa forma, a Tabela 4 apresenta os principais resultados obtidos com o ajuste dos modelos propostos.

Modelo	Parâmetro	Estimativa	Erro padrão	P-valor
Beta	β_0	0,215	0,109	0,049
	β_1	0,016	0,003	<0.001
Simplex	β_0	0,233	0,120	0,052
	β_1	0,015	0,003	<0.001
Normal	β_0	0,564	0,025	<0.001
	β_1	0,003	0,001	<0.001

Tabela 4 – Resultados dos modelos ajustados para o aproveitamento.

Com um nível de significância de 5%, o intercepto apresentou-se significativo para os preditores Beta e Normal. Além disso, o saldo de gols foi significativo para as três variáveis resposta. Dessa forma, os modelos podem ser reformulados da seguinte maneira:

$$\eta_B = 0{,}215 + 0{,}016Gd,$$
$$\eta_S = 0{,}015Gd,$$
$$\eta_N = 0{,}564 + 0{,}003Gd.$$

Os resultados mostram que existe uma relação positiva entre o saldo de gols e o aproveitamento médio dos campeões do Brasileirão. No modelo Beta, que utiliza a transformação logito para interpretar a média, definida em (2.5), observa-se que mesmo quando o saldo de gols é zero, o aproveitamento médio dos campeões já é de cerca de 55%, aumentando em torno de 1,6% a cada gol extra no saldo. O modelo Simplex apresenta comportamento semelhante, partindo de uma condição neutra, e confirma que cada gol adicional eleva o aproveitamento médio em aproximadamente 1,5%, mostrando que maior saldo de gols está associado a desempenho superior. Já o modelo Normal oferece uma leitura direta: com saldo zero, o aproveitamento esperado é de 56,4%, e cada gol extra acrescenta cerca de 0,3 ponto percentual, evidenciando de forma clara como a diferença de gols impacta o desempenho médio dos campeões. Os dados mostram que os campeões costumam ter aproveitamento acima do previsto pelos modelos, variando entre 58% e 79%, como o Cruzeiro em 2003 (72,5%) e o Flamengo em 2019 (78,9%), indicando que, além do saldo de gols, outros fatores podem estar influenciando o desempenho, embora nenhuma outra variável fora estatisticamente significante.

Após o ajuste dos três modelos para o aproveitamento, tendo o saldo de gols como variável preditora, será realizada a avaliação da qualidade do ajuste, com ênfase na análise dos resíduos. O objetivo é verificar se os resíduos seguem, de forma aproximada, a distribuição Normal. A Figura 16 mostra que a maioria dos pontos se alinham bem à reta de referência da Normalidade, havendo apenas algumas exceções nas extremidades. Essas discrepâncias, no entanto, não parecem comprometer o ajuste como um todo, de modo que os modelos podem ser considerados adequados para descrever a relação entre saldo de gols e aproveitamento dos campeões do Brasileirão.

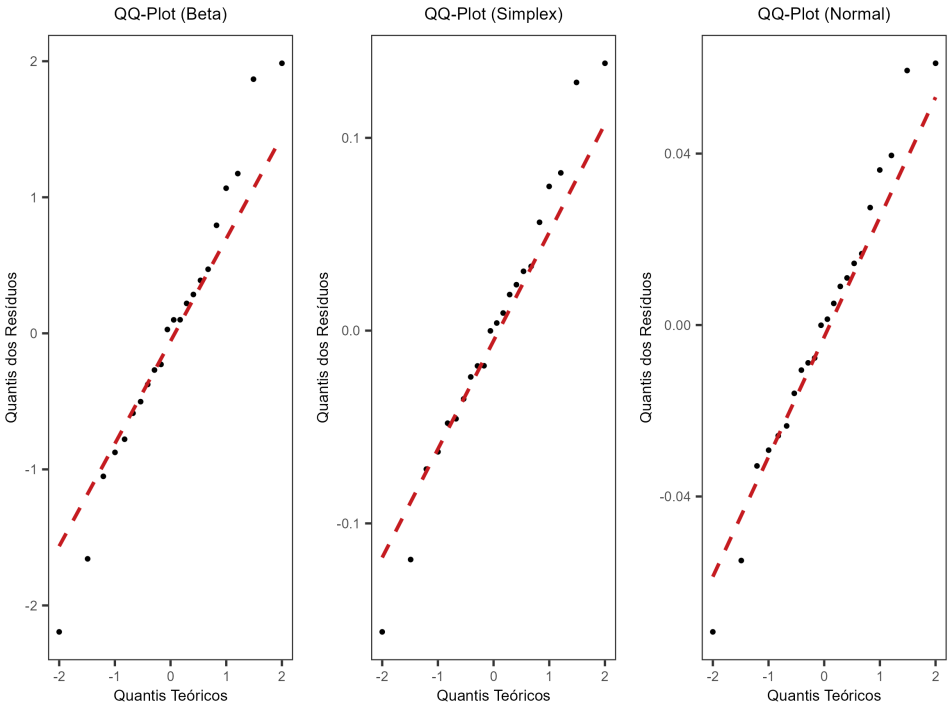

Figura 16 – Verificação da qualidade do ajuste do aproveitamento.

Fonte: Elaboração do autor.

A Figura 18 apresenta a comparação entre os valores observados de aproveitamento ao longo dos anos e os valores ajustados pelos três modelos. O objetivo é avaliar o quão bem cada modelo consegue captar a tendência e a variação dos dados reais ao longo do tempo.

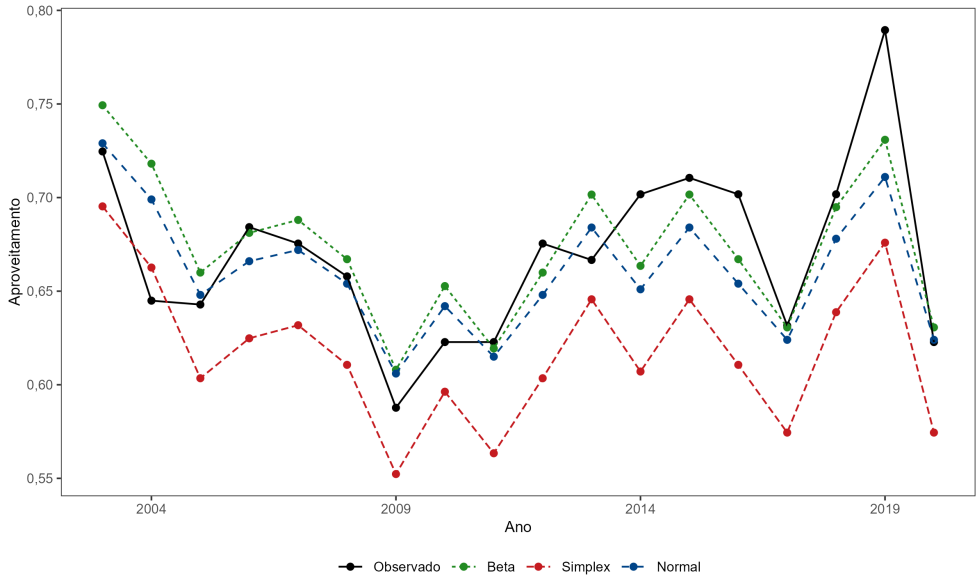

Figura 17 – Aproveitamento observado e estimado pelos modelos.

Fonte: Elaboração do autor.

Observa-se que o modelo Normal apresenta a trajetória mais próxima dos valores observados de aproveitamento ao longo dos anos, reproduzindo melhor as oscilações da série. Embora o modelo Beta também acompanhe relativamente bem os dados, ele se distancia mais em certos pontos. O modelo Simplex, por sua vez, tende a subestimar os valores observados, especialmente nos anos com maior aproveitamento, evidenciando um ajuste inferior. Essa subestimação pode estar ocorrendo porque o intercepto do modelo Simplex não apresentou significância estatística no ajuste. Por esse motivo, decidiu-se manter o intercepto fixo em zero, o que pode reduzir a capacidade do modelo de acompanhar os valores mais elevados ao longo dos anos. No entanto, a análise foi conduzida dessa forma, uma vez que a avaliação do modelo se baseia nos valores-p.

A Tabela 5 mostra os valores de AIC e BIC dos modelos, usados para comparar a qualidade dos ajustes. Valores menores indicam modelos mais adequados, com bom equilíbrio entre ajuste e complexidade.

Modelo	AIC	BIC
Beta	-81,849	-78,576
Simplex	71,776	75,050
Normal	-82,187	-78,914

Tabela 5 – Valores dos critérios AIC e BIC para os modelos ajustados.

Observa-se que o modelo Normal apresenta os menores valores de AIC e BIC, seguido de perto pelo modelo Beta. Já o modelo Simplex teve os maiores valores para ambos os critérios.

4.1.1 Modelo Preditivo

Para a construção do modelo preditivo, os dados foram particionados aleatoriamente em dois subconjuntos: treino e teste. Essa estratégia possibilita estimar os parâmetros com parte das observações e, em seguida, verificar o desempenho do modelo em dados não utilizados no ajuste, oferecendo uma medida mais fiel de sua capacidade de generalização. No total, foram utilizadas 17 observações para o conjunto de treino e 5 para o teste. Embora a amostra contenha apenas 22 casos, o procedimento foi mantido, pois o foco da análise é compreender o desempenho do modelo mesmo em contextos com tamanho amostral reduzido. Essa abordagem é utilizada para examinar tanto a precisão das previsões quanto a adequação dos modelos selecionados.

Para examinar a relação entre o aproveitamento de pontos e o saldo de gols, foram utilizados modelos preditivos com base nas distribuições Beta, Simplex e Normal. A estimação dos parâmetros foi realizada a partir das observações do conjunto de treino, considerando o saldo de gols como variável explicativa e o aproveitamento como resposta. Em seguida, os coeficientes obtidos foram aplicados às observações do conjunto de teste,

com o objetivo de gerar previsões fora da amostra de ajuste. As estimativas foram calculadas manualmente com base nos dados do teste, por meio da função inversa do logit para os modelos Beta e Simplex, e da função identidade para o modelo Normal. Essa etapa é utilizada para avaliar o desempenho dos modelos em dados não utilizados no processo de estimação.

A partir disso, são apresentados os principais resultados obtidos com o ajuste dos modelos propostos.

Modelo	Parâmetro	Estimativa	Erro padrão	P-valor
Beta	β_0	0,319	0,115	0,005
	β_1	0,013	0,003	<0.001
Simplex	β_0	0,333	0,121	0,006
	β_1	0,012	0,004	<0.001
Normal	β_0	0,586	0,028	<0.001
	β_1	0,003	0,001	<0.001

Tabela 6 – Resultados dos modelos preditivos na modelagem do aproveitamento.

Com um nível de significância de 5%, o intercepto e o saldo de gols apresentaram-se significativos para as três variáveis resposta. Dessa forma, os modelos podem ser reformulados da seguinte maneira:

$$\eta_B = 0{,}319 + 0{,}013Gd,$$
$$\eta_S = 0{,}333 + 0{,}012Gd,$$
$$\eta_N = 0{,}586 + 0{,}003Gd.$$

Os modelos preditivos confirmam que quanto maior o saldo de gols, maior o aproveitamento dos campeões. Por (2.5), o aproveitamento médio no modelo Beta parte de cerca de 66% com saldo zero, aumentando 1,3 ponto percentual a cada gol. No Simplex, começa em 33,3%, crescendo 1,2 ponto percentual por gol extra. No modelo Normal, inicia em 58,6%, com acréscimo de 0,3 ponto percentual por gol adicional. Apesar das diferenças nos valores iniciais e nos incrementos, todos os modelos mostram que times com saldo positivo apresentam médias de aproveitamento mais altas. Além disso, mesmo sem gols extras, os campeões já apresentam um nível mínimo de desempenho esperado. Os resultados evidenciam que o efeito do saldo de gols sobre o aproveitamento é consistente entre diferentes abordagens de modelagem. Dessa forma, é possível observar como a vantagem ofensiva tende a se refletir nas médias de desempenho final.

Após o ajuste com os dados do conjunto de treino, a ideia é avaliar a capacidade preditiva dos modelos, verificando a qualidade das previsões realizadas no conjunto de teste, ou seja, fora da amostra utilizada para estimação. Os resultados mostram que os três modelos reproduzem de forma consistente a tendência geral observada ao longo dos anos, acompanhando adequadamente a trajetória real dos dados.

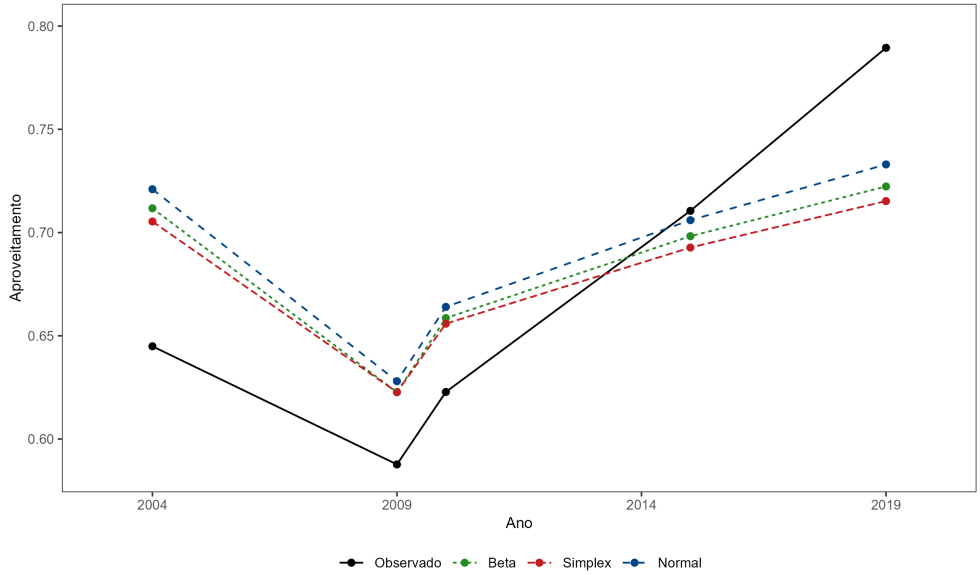

Figura 18 – Predições dos modelos e aproveitamento observado em dados de teste.
Fonte: Elaboração do autor.

Outro método utilizado para avaliar a capacidade preditiva dos modelos ajustados aos dados do Campeonato Brasileiro foi a validação cruzada, que consiste em ajustar o modelo repetidamente, retirando uma única observação por vez do conjunto de dados. A cada iteração, os dados remanescentes foram utilizados para treinar os modelos, enquanto a observação excluída serviu como conjunto de teste. Em cada rodada, foram ajustados os três modelos estudados, utilizando as respectivas funções de ligação previamente definidas: logit para os modelos Beta e Simplex, e identidade para o modelo Normal.

Após o ajuste dos modelos, foi construída uma matriz contendo os valores da variável explicativa da observação deixada de fora, referente ao saldo de gols, que serviu como base para a predição da variável resposta do aproveitamento de pontos. Para cada modelo, foram extraídos os coeficientes estimados e seus respectivos p-valores. Apenas os coeficientes estatisticamente significativos ao nível de 5% foram mantidos nas equações preditivas, enquanto os demais foram zerados, com o objetivo de eliminar o efeito de variáveis não significativas nas previsões.

Com os coeficientes selecionados, foram realizadas as predições para a observação excluída. No caso dos modelos Beta e Simplex, como utilizam a ligação logit, a predição foi obtida aplicando a transformação logística ao produto entre a matriz de teste e os coeficientes estimados. Já para o modelo Normal, a previsão foi feita diretamente por meio de uma combinação linear dos coeficientes. As previsões obtidas em cada iteração foram comparadas ao valor real da variável de resposta da observação deixada de fora, e os erros foram calculados por duas métricas: o viés absoluto e o erro quadrático médio. Importante destacar que esses erros foram computados separadamente para cada um dos três modelos ao longo de todas as iterações, sem que um modelo fosse previamente escolhido. Ao final

das N iterações (onde N representa o número total de observações), foram calculadas as médias dos erros obtidos em cada modelo, possibilitando a comparação de seu desempenho preditivo. A Tabela 7 apresenta estes resultados.

Modelo	|Viés|	EQM
Beta	0.038	0.002
Simplex	0.050	0.003
Normal	0.028	0.001

Tabela 7 – Métricas de erro dos modelos preditivos obtidas via validação cruzada.

Os resultados mostraram que o modelo Normal apresentou menores erros em comparação aos demais modelos, indicando desempenho superior na tarefa de previsão. A validação cruzada não mostra indício de overfitting em nenhum dos modelos, uma vez que os erros se mantêm baixos e próximos entre si.

5 CONCLUSÃO

Neste trabalho, foi possível explorar o processo de análise de regressão aplicado a dados restritos ao intervalo unitário. O aprofundamento nos modelos Beta, Simplex e Normal revelou-se valioso tanto para a compreensão teórica quanto para a realização das simulações. Apesar da relevância desses modelos, alguns desafios e limitações foram encontrados ao longo da análise. Um dos principais obstáculos refere-se à distribuição Normal, que não permite, de forma nativa, delimitar a geração de valores aleatórios dentro do intervalo de 0 a 1, diferentemente das distribuições Beta e Simplex, que já operam naturalmente nesse domínio. Na simulação realizada com a distribuição Normal, por exemplo, os valores gerados oscilaram entre aproximadamente 0,883 e 2,234, com média próxima de 1,520, ultrapassando, portanto, o limite superior esperado para distribuições unitárias. Esse comportamento evidencia um obstáculo do modelo Normal na simulação, ainda que, em termos de estimativas, o ajuste tenha apresentado um desempenho satisfatório. Outro ponto a ser considerado é que tanto nas simulações quanto na aplicação prática foi utilizado um número reduzido de variáveis independentes e isso pode ter limitado, em certa medida, a possibilidade de generalização dos resultados e a exploração do desempenho dos métodos em contextos mais complexos, com múltiplos preditores.

Inicialmente, explorou-se a relevância e as especificidades envolvidas na análise de dados restritos ao intervalo unitário e, em seguida, foram abordadas as distribuições Beta, Simplex e Normal, suas respectivas parametrizações e a forma como podem ser utilizadas na construção de modelos de regressão. A partir das propriedades, transformações e visualizações discutidas, torna-se viável realizar uma abordagem inicial desse tipo de dado e das respectivas modelagens.

Com a teoria previamente apresentada, as simulações permitiram observar um bom desempenho dos estimadores dos coeficientes de regressão. Notou-se uma redução progressiva do erro padrão, do viés absoluto e do erro quadrático médio à medida que o tamanho das amostras aumentava. Destaca-se, nesse contexto, o desempenho superior do modelo Normal em relação às distribuições unitárias, evidenciado pelos menores valores obtidos para essas medidas de erro.

Nas aplicações realizadas, foi possível empregar as técnicas previamente apresentadas na construção dos modelos de regressão. Para a análise dos dados dos Campeões Brasileiros, o modelo Normal apresentou bom desempenho, se destacando em diferentes aspectos e mostrando melhor ajuste em comparação aos demais modelos. Apesar do suporte real, o modelo Normal mostrou-se competitivo frente a alternativas específicas para dados no intervalo unitário.

REFERÊNCIAS

ANDRADE, Augusto César Giovanetti de. Efeitos da especificação incorreta da função de ligação no modelo de regressão beta. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2007.

AKRAM, Saba; QURRAT UL ANN. Newton-Raphson Method. International Journal of Scientific Engineering Research, v. 6, n. 7, p. 1748–1752, 2015.

BARNDORFF-NIELSEN; JORGENSEN, B. Some parametric models on the simplex. Journal of Multivariate Analysis, Elsevier, v. 39, n. 1, p. 106-116, 1991.

CRIBARI-NETO, F.; ZEILEIS, A. "Beta Regression in R." Journal of Statistical Software, v. 34, n. 2, p. 1–24, 2010.

CRIBARI-NETO, F.; FERRARI, S. Beta regression for modeling rates and proportions. Journal of Applied Statistics, v. 31, p. 799-815, 2004.

DONATELLI, G.D.; KONRATH, A.C. Simulação de Monte Carlo na Avaliação de Incertezas de Medição. Revista de Ciência & Tecnologia, v. 13, n. 25/26, p. 5-15, 2005.

FERNANDES, F. H.; SILVA, R. S. Distribuição Simplex: Avaliação e Estimação dos Parâmetros. Revista Brasileira de Estatística, v.77, n. 242, p. 33-51, 2019.

Galton, F. Natural Inheritance. London: Macmillan. 1889.

JORGENSEN, B. The Theory of Dispersion Models, Technometrics, v. 41, n. 2, p. 177–178, 1997.

JÚNIOR, A. F. S.; TABOSA, C. M.; COSTA, R. P. Simulação de Monte Carlo aplicada à análise econômica de pedido. Produção, v. 21, n. 1, p. 149-164, 2011.

KIESCHNICK, R.; MCCULLOUGH, BD. "Regression Analysis of Variates Observed on (0,1): Percentages, Proportions and Fractions." Statistical Modelling, v. 3, n. 3, p. 193–213, 2003.

LIMA, Francimário Alves de. Distribuições de probabilidade no intervalo unitário. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

MAGALHÃES, T. M.; GALLARDO, D. I.; DINIZ, M. A. Inferential and predictive procedures for inverse gamma regression model. Journal of Statistical Computation and Simulation, v. 95, n. 11, p. 2327-2342, 2025.

MCCULLAGH, P.; NELDER, J. A. Generalized Linear Models. Chapman and Hall, London, 1989.

MOONEY, C. Z. Monte Carlo simulation Thousand Oaks: Sage Publications, 1997.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. Applied Linear Statistical Models, Chicago: Irwin, 4th ed., 1996.

OLIVEIRA, Marcos Santos de.; Um modelo de regressão beta: teoria e aplicações. Dissertação de Mestrado, Universidade de São Paulo, São Paulo, 2004.

SILVA, A. de O. Regressão Simplex não Linear: Inferência e Diagnóstico. Dissertação de Mestrado, Universidade Federal de Pernambuco, 2015.

SIMAS, A. B; BARRETO-SOUZA, W.; ROCHA, A. V. "Improved Estimators for a General Class of Beta Regression Models." Computational Statistics & Data Analysis, v. 54, ed. 2, p. 348–366, 2010.

SIMAS, A. B.; ROCHA, A. V. betareg: Beta Regression. R package version 1.2, 2006. Disponível em: `https://CRAN.R-project.org/src/contrib/Archive/betareg/`.

SMITHSON, M.; VERKUILEN, J. "A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables." Psychological Methods, v. 11, n. 1, p. 54–71, 2006.

TRUNFIO, T.A., SCALA, A., GIGLIO, C. et al. Multiple regression model to analyze the total LOS for patients undergoing laparoscopic appendectomy. BMC Medical Informatics and Decision Making, v. 22, n. 141, 2022.

VAN DEN BOOGAART, K. G.; TOLKSDORF, M.; TSCHIRREN, L.; LEHMANN, R.; NEUHAUS, F. compositions: Compositional Data Analysis. R package version 2.0-6. Disponível em:
`https://cran.r-project.org/web/packages/compositions/index.html`.

ZALUSKA, M.; GLADYSZEWSKA-FIEDORUK, K. Regression Linear Model of Air Pollution Emission on the Example of a Waste Incineration Plant. Proceedings, v. 51, n. 32, 2020.

ZERBINATTI, L. F. M.; FERRARI, S. L. d. P. Predição de fator de simultaneidade através de modelos de regressão para proporções contínuas. Dissertação de Mestrado, Universidade de São Paulo, São Paulo, 2008.

ZHANG, P.; QIU, Z.; SHI, C. Simplexreg: An R Package for Regres sion Analysis of Proportional Data Using the Simplex Distribution. Journal of Statistical Software, v. 71, n. 11, p. 1-24, 2016. Disponível em:
`https://cran.r-project.org/web/packages/simplexreg/index.html`.