Universidade de Brasília
Departamento de Estatística

Estudo sobre modelos de regressão beta

Eduardo de Sousa Carvalho

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Eduardo de Sousa Carvalho

Estudo sobre modelos de regressão beta

Orientadora: Profa. Terezinha Késsia de Assis Ribeiro

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Resumo

A distribuição beta é comumente utilizada para modelar dados contínuos limitados ao intervalo unitário. Esta possui grande flexibilidade, a depender dos valores assumidos por seus parâmetros, podendo apresentar diversas formas. A reparametrização da distribuição beta viabilizou a utilização desta no contexto de modelos de regressão, os quais apresentam grande potencial para ajustar dados limitados ao intervalo (0,1) produzidos por diversos fenômenos aleatórios. As distribuições beta inflacionadas foram desenvolvidas com o objetivo de agregar os valores 0 e/ou 1 à distribuição beta original e, desse modo, suprir uma limitação daquele modelo, uma vez que dados limitados ao intervalo unitário podem, por vezes, assumir os valores 0 ou 1. Neste trabalho, discutimos as principais características das distribuições beta e beta inflacionadas, além dos respectivos modelos de regressão desenvolvidos a partir destas distribuições de probabilidade. Para ilustrar a aplicabilidade destes modelos, foram efetuados estudos de simulação e aplicações a dados reais por meio dos quais foi possível identificar as vantagens e limitações de cada método.

Palavras-chave: Distribuição beta, Distribuição beta inflacionada, Modelos de regressão beta, Modelos de regressão beta inflacionados.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A modelagem adequada de dados contínuos limitados ao intervalo unitário surge naturalmente como um obstáculo a ser superado em diversas áreas do conhecimento. Tais tipos de dados são usualmente taxas, proporções, percentagens e frações. Alguns exemplos são fração da renda familiar gasta com alimentação, escores de qualidade de vida, proporção do tempo que animais gastam com uma atividade, e percentual da superfície coberta pela vegetação em uma região.

Para lidar com dados que possuem tal característica dentro do contexto de regressão, pode-se modelar a média μ_t de uma variável y_t denominada de resposta que assume valores no intervalo (0,1) em função de outras variáveis que são conhecidas e fixadas. Estas últimas são comumente chamadas de covariáveis ou variáveis explicativas. Uma alternativa para modelar a média μ_t é através da técnica de regressão linear normal (DRAPER; SMITH, 1998). Nesta abordagem, a variável resposta y_t segue uma distribuição normal com média μ_t e variância constante σ^2. Supõe-se que a média μ_t se relaciona com k covariáveis através da estrutura $\mu_t = X_t^\top \beta$ com $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^\top \in \mathbb{R}^k$ sendo um vetor de parâmetros desconhecido que deve ser estimado, e $X_t = (x_{t1}, x_{t2}, \ldots, x_{tk})^\top \in \mathbb{R}^k$ o vetor com os valores das k covariáveis para a t-ésima observação da amostra. Entretanto, o uso desta abordagem pode conduzir a alguns problemas relacionados com a inferência de quantidades desconhecidas do modelo.

Primeiro, uma forma de ajustar os valores y_t através desta abordagem é através de $\widehat{y}_t = \widehat{\mu}_t = X_t^\top \widehat{\beta} \in \mathbb{R}$, ou seja, $\widehat{y}_t$ assume valores reais. Dessa forma, ao ajustar y_t através de $\widehat{y}_t$ podemos obter valores que não pertencem ao intervalo contínuo unitário. O mesmo problema pode ocorrer ao realizar previsões para valores de y fixados valores de X que não foram observados na amostra. Segundo, supõe-se que a distribuição de y_t é simétrica (normal) com variância constante, ou seja, o modelo probabilístico assumido para y_t é simétrico em torno de sua média μ_t e homocedástico. Entretanto, dados contínuos limitados ao intervalo (c,d), com c e d constantes, podem possuir distribuição assimétrica e variância não constante (QUEIROZ, 2022). Tais características podem conduzir altos vieses nas estimativas dos parâmetros desconhecidos do modelo, e consequentemente, produzir um ajuste inadequado aos dados. De forma geral, o ajuste de um modelo de regressão linear normal para dados limitados ao intervalo unitário não é uma abordagem adequada.

Para dados desta natureza, dentro do contexto de modelos de regressão, se torna apropriado supor uma distribuição de probabilidades para y_t que tenha suporte no intervalo (0,1), acomode diversas formas, e que possua variância dependente da média μ_t. Na literatura existem algumas propostas para lidar com estes dados que são basea-

das em distribuições de probabilidades com suporte no (0,1) (BARNDORFF-NIELSEN; JØRGENSEN, 1991; KIESCHNICK; MCCULLOUGH, 2003; SONG; TAN, 2000; FERRARI; CRIBARI-NETO, 2004). Para o desenvolvimento deste trabalho, focaremos em abordagens em que a variável resposta y_t segue uma distribuição de probabilidades beta ou uma mistura de distribuições que envolve a distribuição beta.

O modelo probabilístico beta é uma distribuição de probabilidades com dois parâmetros associado a uma variável aleatória contínua que assume valores no intervalo (0,1). A depender da combinação de seus dois parâmetros, esta distribuição assume diversas formas, incluindo formas assimétricas. Considerando uma reparametrização desta distribuição, Ferrari e Cribari-Neto (2004) propuseram uma classe de modelos de regressão em que y_t segue uma distribuição beta indexada pela média μ_t e por um parâmetro de precisão ϕ. Nesta abordagem, a média μ_t é modelada através de estrutura de regressão linear $g_\mu(\mu_t) = X_t^\top \beta$ com $g_\mu(\cdot) : (0,1) \to \mathbb{R}$ denominada de função de ligação. Sendo assim, obtém-se que $\mu_t = g_\mu^{-1}(X_t^\top \beta) \in (0,1)$. Logo, ao ajustar y_t por $\widehat{\mu}_t$, sempre será obtido um valor ajustado dentro do intervalo (0,1). Também, este modelo é heterocedástico pois a variância de y_t varia com as observações através de sua média μ_t. Uma extensão natural desta abordagem é supor que a precisão dos dados também varie de acordo com as observações. Tal proposta foi introduzida por Smithson e Verkuilen (2006) onde supõe-se que y_t segue uma distribuição beta indexada pela média μ_t e precisão ϕ_t. Nesse sentido, atribui-se uma estrutura de regressão linear para a precisão ϕ_t.

As duas últimas propostas se mostram adequadas para dados contínuos que estão limitados no intervalo (0,1). Entretanto, na prática, podemos nos deparar com a ocorrência de zeros e/ou uns, isto é, observar fenômenos aleatórios que produzem dados nos intervalos [0,1), (0,1] ou [0,1]. Uma forma de resolver este problema é aplicar uma transformação nas observações iguais a 0 ou 1 de forma que o valor resultante esteja em $(0,1)$, e assim, utilizar as abordagens discutidas anteriormente. Entretanto, ao realizar este procedimento, pode-se introduzir viés severo nas estimativas dos parâmetros, e assim conduzir a conclusões erradas sobre as características de interesse. Na verdade, ao fazer este tipo de procedimento, o viés introduzido em estimativas de máxima verossimilhança não será limitado (RIBEIRO; FERRARI, 2022).

Uma forma adequada de lidar com tais situações é fazer uso da classe de modelos de regressão beta inflacionados introduzida por Ospina (2008). Nesta proposta são introduzidos dois novos métodos de modelagem baseados na distribuição beta e que contemplam a ocorrência de zeros, uns, ou ambos os valores extremos. Na primeira abordagem supõe-se que a variável y_t segue uma distribuição beta inflacionada no ponto c, que é definida como sendo uma mistura entre a distribuição beta contínua e uma distribuição discreta degenerada no ponto $y = c$ com c sendo igual a zero ou um. A segunda abordagem parte do pressuposto de que y_t é distribuída segundo uma mistura entre a

distribuição beta contínua e uma distribuição de *Bernoulli*.

O presente Relatório final, que está organizado em nove capítulos, se propõe a estudar as classes de modelos de regressão beta introduzidas por Ferrari e Cribari-Neto (2004), Smithson e Verkuilen (2006) e Ospina (2008), além de realizar comparações entre estes modelos através das suas especificações. No Capítulo 2 são descritos os modelos probabilísticos estudados e as suas principais características. No Capítulo 3 são apresentados os modelos de regressão beta desenvolvidos a partir de cada distribuição apresentada no Capítulo 2. No Capítulo 4 são discutidos métodos para obtenção de estimativas intervalares e para realização de testes de hipóteses sobre os parâmetros dos modelos de regressão. No Capítulo 5 são apresentadas algumas medidas de informação que podem ser utilizadas como critérios para seleção de modelos. No sexto capítulo são discutidas técnicas para diagnóstico aplicáveis aos modelos de regressão beta. O Capítulo 7 apresenta a metodologia utilizada nos estudos de simulação e nas aplicações, além de especificar os recursos computacionais e softwares utilizados. No Capítulo 8 são efetuados os estudos de simulação e as aplicações a dados reais. Por fim, no Capítulo 9 são apresentadas as considerações finais deste trabalho.

2 Distribuições de probabilidade

2.1 Distribuição beta

A distribuição beta é uma família de distribuições de probabilidade contínuas definida com suporte no intervalo (0,1) e parametrizada por dois elementos, ambos positivos, aqui denotados por a e b (CASELLA; BERGER, 2011). Esses dois parâmetros aparecem como expoentes na função densidade da variável aleatória e controlam a forma da distribuição.

A função densidade de probabilidade (fdp) de uma variável aleatória y que segue uma distribuição beta de parâmetros a, $b > 0$ é definida por

$$f(y; a,b) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}, \quad 0 < y < 1, \tag{2.1.1}$$

em que $B(a,b)$ é a função beta dada por

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

e $\Gamma(.)$ é a função gama definida por

$$\Gamma(z) = \int_0^\infty u^{z-1}e^{-u}du,$$

sendo z um número complexo cuja parte real é estritamente positiva. Assim, a expressão (2.1.1) pode ser reescrita como

$$f(y; a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}y^{a-1}(1-y)^{b-1}, \quad 0 < y < 1 \text{ e } a, b > 0. \tag{2.1.2}$$

A função de distribuição acumulada (fda) da distribuição beta é definida por

$$\begin{aligned} F(y; a,b) &= \int_{-\infty}^{y} f(t; a,b)dt \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^y t^{a-1}(1-t)^{b-1}dt \\ &= \frac{1}{B(a,b)}\int_0^y t^{a-1}(1-t)^{b-1}dt \\ &= \frac{B_y(a,b)}{B(a,b)}, \end{aligned} \tag{2.1.3}$$

em que $0 < y < 1$ e $B_Y(a,b) = \int_0^y t^{a-1}(1-t)^{b-1}dt$ é conhecida por função beta incompleta

(JOHNSON; KOTZ; BALAKRISHNAN, 1995).

A distribuição beta é bastante flexível a depender dos valores assumidos pelos parâmetros a e b. Esta pode exibir uma infinidade de formas, sendo amplamente utilizada pra modelar o comportamento de diversos tipos de fenômenos aleatórios, desde que a variável de interesse assuma valores limitados ao intervalo (0,1). Não obstante, a distribuição beta é também aplicável a fenômenos que produzem valores no intervalo (c, d), com c e d constantes reais. Para tanto, aplica-se a transformação $(y - c)/(d - c)$ para representar esse intervalo contínuo dentro do suporte exigido para a distribuição beta.

Adotando a função densidade de probabilidade em (2.1.2), os momentos de ordem n, com $n = 1,2,3,\ldots$, podem ser obtidos diretamente pela definição

$$\mathrm{E}(y^n) = \int_0^1 y^n f(y; a,b)dy.$$

Assim, temos que

$$\begin{aligned}
\mathrm{E}(y^n) &= \frac{1}{B(a,b)} \int_0^1 y^{(a+n)-1}(1-y)^{b-1}dy \\
&= \frac{B(a+n,b)}{B(a,b)} \int_0^1 \frac{1}{B(a+n,b)} y^{(a+n)-1}(1-y)^{b-1}dy \\
&= \frac{B(a+n,b)}{B(a,b)} \\
&= \frac{\Gamma(a+n)\Gamma(b)}{\Gamma(a+n+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}.
\end{aligned}$$

Simplificando, obtemos que

$$\mathrm{E}(y^n) = \frac{\Gamma(a+n)}{\Gamma(a+n+b)} \frac{\Gamma(a+b)}{\Gamma(a)}.$$

Tomando $n = 1,2$, e usando propriedades da função gama, obtemos os dois primeiros momentos de y, por meio dos quais chegamos a expressões fechadas para, respectivamente, a média e a variância da variável aleatória y. Portanto,

$$\mathrm{E}(y) = \frac{\Gamma(a+1)}{\Gamma(a+1+b)} \frac{\Gamma(a+b)}{\Gamma(a)} = \frac{a}{a+b},$$

$$\begin{aligned}
\mathrm{Var}(y) &= \mathrm{E}(y^2) - [\mathrm{E}(y)]^2 \\
&= \frac{\Gamma(a+2)}{\Gamma(a+2+b)} \frac{\Gamma(a+b)}{\Gamma(a)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{ab}{(a+b)^2(a+b+1)}.
\end{aligned}$$

Ferrari e Cribari-Neto (2004) propuseram uma reparametrização da distribuição beta reescrevendo (2.1.2) por meio de novos parâmetros que representam a média e a precisão de y. Tal alteração objetivou definir uma estrutura de regressão para modelar a média μ de uma variável resposta y que seja distribuída segundo uma distribuição beta. Além disso, para viabilizar a modelagem da média μ_t foi necessário estabelecer um parâmetro ϕ que representasse a precisão.

Nesse sentido, foi definido $\mu = \mathrm{E}(y) = a/(a+b)$ e $\phi = a+b$, resultando em $a = \mu\phi$ e $b = \phi - \mu\phi = (1-\mu)\phi$ e, consequentemente, na seguinte expressão para a fdp da distribuição beta reparametrizada:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \;\; 0<y<1. \tag{2.1.4}$$

A fda da distribuição beta reparametrizada é da forma

$$F(y;\mu,\phi) = \int_{-\infty}^{y} f(t;\mu,\phi)dt = \frac{B_Y(\mu\phi,(1-\mu)\phi)}{B(\mu\phi,(1-\mu)\phi)}. \tag{2.1.5}$$

Denotaremos por $y \sim \mathcal{B}(\mu,\phi)$ uma variável aleatória y que possui distribuição beta com função densidade de probabilidade na forma (2.1.4). Conforme mencionado anteriormente, a distribuição beta é bastante flexível, resultando em grande potencial para modelar dados limitados ao intervalo (0,1). Na Figura 1 são apresentadas curvas da função densidade de probabilidade da distribuição beta, considerando diferentes valores para os parâmetros μ e ϕ. Percebe-se que as curvas podem apresentar diferentes formas a depender dos valores assumidos pelos parâmetros. Quando $\mu = 0{,}5$ e $\phi \neq 2$, as curvas apresentam formas simétricas e unimodais. Para $\mu \neq 0{,}5$ as formas apresentadas são assimétricas podendo ser unimodais, em formas de J ou J invertido. Para $\mu = 0{,}5$ e $\phi < 2$, a curva assume a forma de U. Quando $\mu = 0{,}5$ e $\phi = 2$, a função densidade da distribuição beta se reduz à da distribuição uniforme padrão.

Sob a nova parametrização, a variância da variável aleatória y passa a ser

$$\begin{aligned}
\mathrm{Var}(y) &= \frac{ab}{(a+b)^2(a+b+1)} \\
&= \frac{\mu\phi(1-\mu)\phi}{(\mu\phi+(1-\mu)\phi)^2(\mu\phi+(1-\mu)\phi+1)} \\
&= \frac{\mu(1-\mu)}{\phi+1} \\
&= \frac{V(\mu)}{\phi+1},
\end{aligned}$$

em que $V(\mu) = \mu(1-\mu)$. Observa-se que, mantendo a média μ constante, a variância

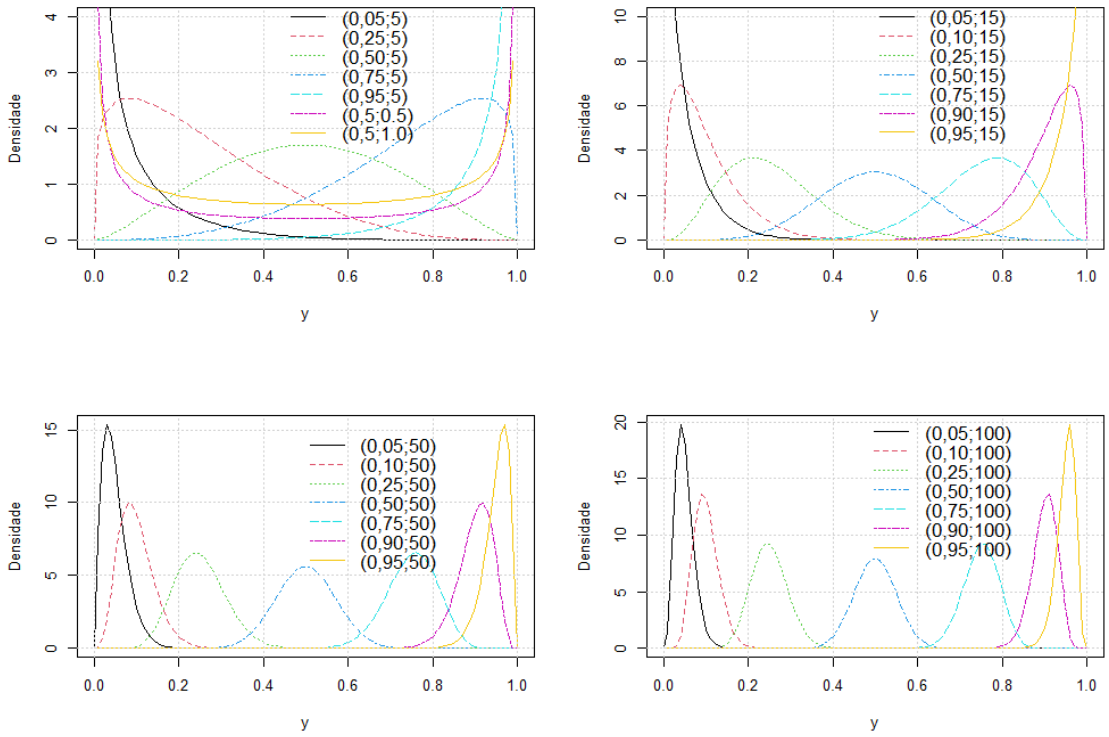

Figura 1 Curvas para a fdp da distribuição beta reparametrizada para diferentes valores de (μ, ϕ).

de y tende a reduzir quanto maior for o valor de ϕ. Em contrapartida, valores baixos de ϕ resultam em valores altos para a variância de y. Por esta razão, ϕ é tido como parâmetro de precisão da distribuição beta reparametrizada. Esse resultado também pode ser visualizado por meio da Figura 2 em que estão representadas algumas curvas para a fdp (2.1.4), com μ fixado em 0,5 e diferentes valores do parâmetro de precisão ϕ.

2.2 Distribuição beta inflacionada

Em situações práticas, dados que representam taxas e proporções podem se concentrar nos extremos do intervalo (0,1) e, por vezes podem incluir também zeros e/ou uns. Nessa situação, os modelos de regressão beta estudados nas seções anteriores não são adequados, uma vez que resultariam em probabilidade zero para valores de y fora do intervalo (0,1).

Para essas situações, Ospina e Ferrari (2010) introduziram a família de distribuições beta inflacionada, objetivando acomodar zeros e/ou uns, e, desse modo, estimar a probabilidade de y assumir um ou ambos os valores extremos. Segundo Ospina (2008), a palavra *inflacionada* sugere que a massa de probabilidade de um ou mais pontos excede a massa de probabilidade permitida sob o modelo original. Na literatura, é comum encon-

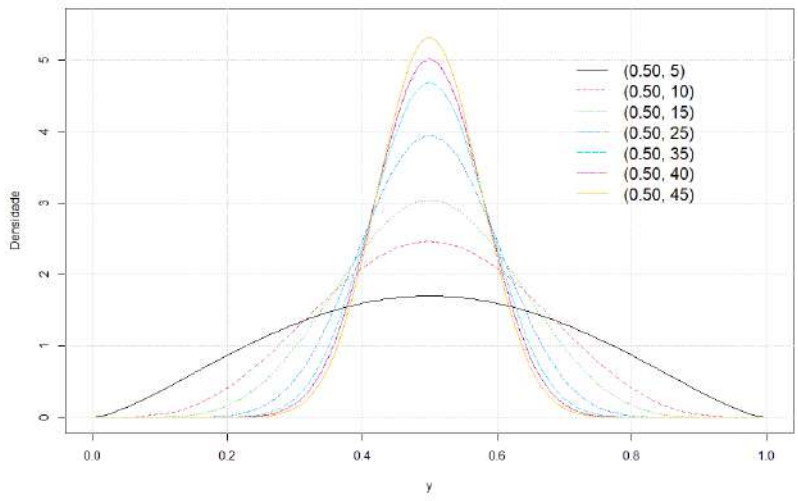

Figura 2 Curvas para a função densidade de probabilidade da distribuição beta reparametrizada para μ fixo e diferentes valores de ϕ.

trarmos também a nomenclatura "aumentada" para se referir a esses tipos de modelos probabilísticos (SILVA, 1989; KODA, 2018).

2.2.1 Distribuição beta inflacionada em zero ou em um

Adequado para dados observados no intervalo [0,1) ou (0,1], o modelo probabilístico conhecido por distribuição beta inflacionada em zero ou em um consiste na mistura entre uma distribuição beta e uma distribuição degenerada em zero ou em um, conforme o caso observado na amostra. Assim, a distribuição beta modela o componente contínuo dos dados, enquanto que a distribuição degenerada ajusta o componente relativo ao ponto de massa em zero ou em um.

Dizemos que uma variável aleatória discreta possui distribuição degenerada quando o seu suporte consiste em um único valor, aqui denotado por c (MACYS, 1987). Segundo Ospina e Ferrari (2010), a fdp da distribuição beta inflacionada em zero ou um (BEZI ou BEOI) é dada por

$$bi_c(y;\alpha,\mu,\phi) = \begin{cases} \alpha, & \text{se } \ y = c \\ (1-\alpha)f(y;\mu,\phi), & \text{se } \ 0 < y < 1, \end{cases} \tag{2.2.1}$$

em que $f(y;\mu,\phi)$ é a fdp da distribuição beta reparametrizada apresentada em (2.1.4), com parâmetros μ e ϕ ($0 < \mu < 1$ e $\phi > 0$), e $0 < \alpha < 1$ representa a probabilidade de observar zero ($c = 0$) ou um ($c = 1$). A função densidade em (2.2.1) também pode ser

representada na forma

$$bi_c(y;\alpha,\mu,\phi) = \{\alpha^{I_{\{c\}}(y)}(1-\alpha)^{1-I_{\{c\}}(y)}\}\{f(y;\mu,\phi)^{1-I_{\{c\}}(y)}\}, \tag{2.2.2}$$

em que $I_{\{c\}}(y)$ representa a função indicadora que assume valor 1 se $y = c$, e 0 caso contrário. Analisando a função densidade em (2.2.2), percebe-se que quando c assume valores diferentes de 0 ou 1, isto é, $y \in (0,1)$, a expressão equivale a (2.1.4) ponderada por $(1-\alpha)$, e quando $c = 0$ ou $c = 1$, assume o valor α.

A fda da mistura é dada por

$$BI_c(y;\alpha,\mu,\phi) = \alpha I_{\{c\}}(y) + (1-\alpha)F(y;\mu,\phi), \tag{2.2.3}$$

em que a função $F(\cdot;\mu,\phi)$ é a fda da distribuição beta em sua forma reparametrizada, conforme (2.1.5).

Conforme definido por Ospina e Ferrari (2010), considerando uma variável aleatória y com a fdp apresentada em (2.2.1), quando $c = 0$, a chamamos de distribuição inflacionada em zero, e denotamos $y \sim \text{BEZI}(\alpha,\mu,\phi)$, e quando $c = 1$, denominamos de distribuição inflacionada em um, e escrevemos $y \sim \text{BEOI}(\alpha,\mu,\phi)$.

A média e variância desta distribuição são expressas por

$$\begin{aligned} \text{E}(y) &= \alpha c + (1-\alpha)\mu, \\ \text{Var}(y) &= (1-\alpha)\frac{V(\mu)}{\phi+1} + \alpha(1-\alpha)(c-\mu)^2, \end{aligned} \tag{2.2.4}$$

em que $V(\mu) = \mu(1-\mu)$. Da expressão (2.2.4), para a distribuição BEZI temos que

$$\begin{aligned} \text{E}(y) &= (1-\alpha)\mu, \\ \text{Var}(y) &= (1-\alpha)\frac{V(\mu)}{\phi+1} + \alpha(1-\alpha)\mu^2, \end{aligned}$$

e para a distribuição BEOI temos

$$\begin{aligned} \text{E}(y) &= \alpha + (1-\alpha)\mu, \\ \text{Var}(y) &= (1-\alpha)\frac{V(\mu)}{\phi+1} + \alpha(1-\alpha)(1-\mu)^2. \end{aligned}$$

Para ambas as distribuições observa-se que para valores fixados de μ e α, a variância de y tende a aumentar conforme o valor de ϕ diminui, o que nos mostra que para estas distribuições ϕ também pode ser interpretado como um parâmetro de precisão.

Nas Figuras 3 e 4 são apresentadas algumas curvas para as fdps das distribuições BEZI e BEOI, respectivamente, considerando diferentes valores para os parâmetros α, μ

e ϕ. Conforme se observa nos gráficos, as curvas mantém a flexibilidade verificada para a distribuição beta com suporte limitado ao intervalo (0,1), a depender da combinação dos valores dos parâmetros μ e ϕ. Entretanto, dada a adição do novo parâmetro α, não é possível obter formas simétricas uma vez que temos a presença de massa de probabilidade no ponto zero ou no ponto um, conforme o caso.

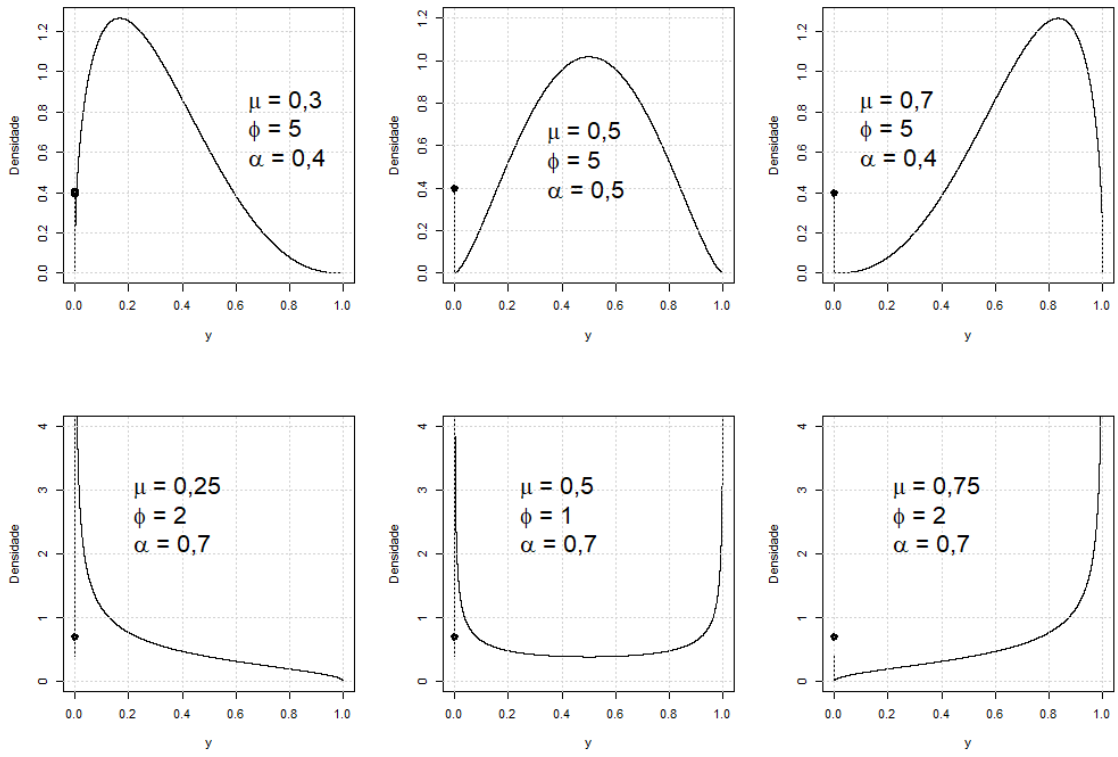

Figura 3 Curvas para a fdp da distribuição BEZI para diferentes valores de ϕ, μ e α.

2.2.2 Distribuição beta inflacionada em zero e um

Quando os dados estão contidos no intervalo contínuo [0,1], ou seja, incluindo ambos os extremos, o modelo beta inflacionado introduzido na seção anterior não é adequado. Para essa situação, foi desenvolvida a distribuição beta inflacionada em zero e um. Esta consiste na mistura entre uma distribuição beta, que modela o componente contínuo da distribuição, e uma distribuição de Bernoulli, para modelar o componente discreto, que são as observações 0 e 1.

A distribuição de *Bernoulli* é um modelo probabilístico discreto baseado na ideia de um experimento aleatório que possui somente dois resultados possíveis, sendo um deles identificado como "sucesso", com probabilidade γ de ocorrência, e o outro como "fracasso", com probabilidade $1 - \gamma$ (ROSS, 2009). Então, dizemos que uma variável

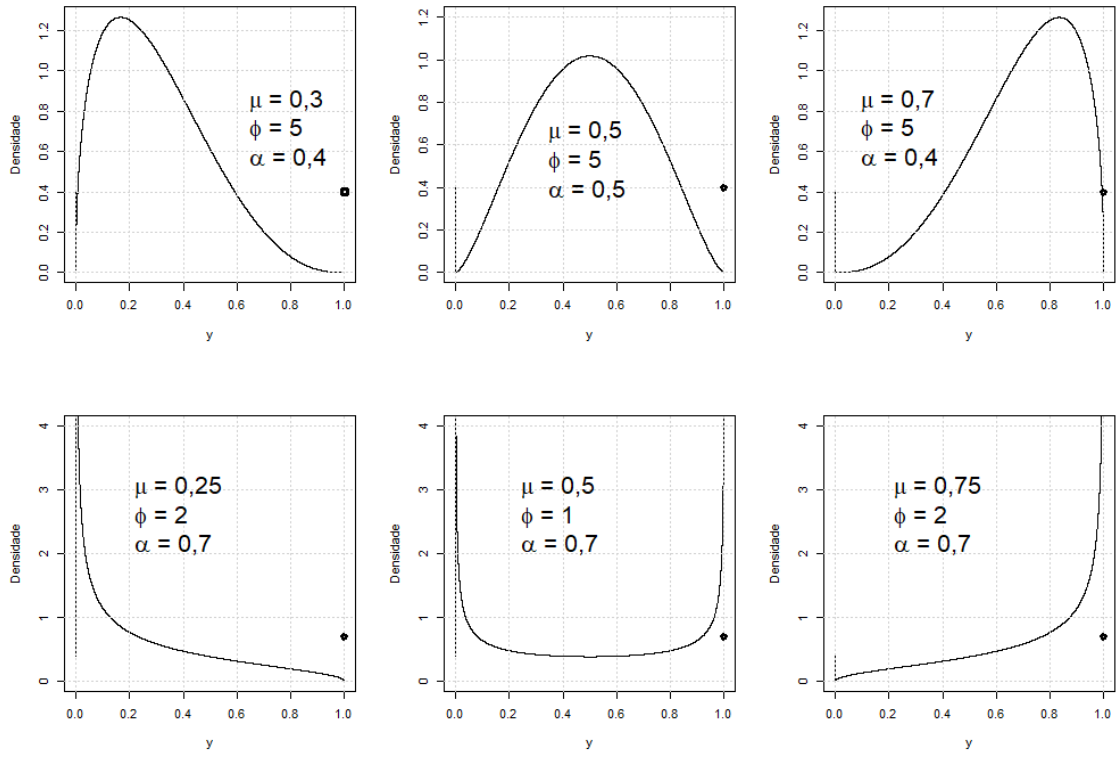

Figura 4 Curvas para a função densidade de probabilidade da distribuição BEOI para diferentes valores de ϕ, μ e α.

aleatória y possui distribuição de *Bernoulli* com parâmetro γ se sua fdp é

$$f(y;\gamma) = \gamma^y(1-\gamma)^{1-y},$$

em que $y = 0{,}1$ e $0 \leq \gamma \leq 1$. Denotamos $y \sim$ *Bernoulli*(γ), com os momentos de ordem n, $n = 1{,}2{,}3, \ldots$, dados por

$$\mathrm{E}(y^n) = 0^n\gamma^0(1-\gamma)^{1-0} + 1^n\gamma^1(1-\gamma)^{1-1} = \gamma,$$

e, portanto,

$$\begin{aligned}\mathrm{E}(y) &= \mathrm{E}(y^1) = \gamma, \text{ e}\\ \mathrm{Var}(y) &= \mathrm{E}(y^2) - [\mathrm{E}(y)]^2 = \gamma - \gamma^2 = \gamma(1-\gamma).\end{aligned}$$

A fda de uma variável aleatória y com distribuição de Bernoulli é da forma

$$\mathrm{Ber}(y;\gamma) = \begin{cases} 0, & \text{para } y \leq 0,\\ 1-\gamma, & \text{para } 0 < y < 1,\\ 1, & \text{para } y \geq 1.\end{cases} \tag{2.2.5}$$

Segundo Ospina e Ferrari (2010), a fdp da distribuição beta inflacionada em zero e um (BEINF) é da forma

$$\text{beinf}(y;\pi,\gamma,\mu,\phi) = \begin{cases} \pi\gamma, & \text{se } \ y = 1, \\ \pi(1-\gamma), & \text{se } \ y = 0, \\ (1-\pi)f(y;\mu,\phi), & \text{se } \ 0 < y < 1, \end{cases} \tag{2.2.6}$$

em que $\pi,\gamma \in (0,1)$ e $f(y;\mu,\phi)$ é a fdp da distribuição beta reparametrizada apresentada em (2.1.4), com parâmetros μ e ϕ ($0 < \mu < 1$ e $\phi > 0$).

Ospina (2008) define a fda da distribuição BEINF como

$$\text{BEINF}(y;\pi,\gamma,\mu,\phi) = \pi\text{Ber}(y;\gamma) + (1-\pi)F(y;\mu,\phi), \tag{2.2.7}$$

em que $\text{Ber}(\cdot,\gamma)$ representa a fda de uma variável aleatória de Bernoulli conforme (2.2.5), e $F(\cdot;\mu,\phi)$ é a função de distribuição acumulada da distribuição beta reparametrizada, conforme (2.1.5).

Observa-se que em (2.2.6) o parâmetro π representa a proporção da mistura, que pondera a probabilidade da variável aleatória assumir valores no intervalo contínuo (0,1) ou nos pontos 0 e 1, permitindo assim combinar as duas distribuições citadas. Note, ainda, que π representa a probabilidade da variável aleatória y ser selecionada da distribuição de *Bernoulli*, e $1-\pi$ representa a probabilidade da variável aleatória y ser selecionada da distribuição beta. Denotaremos por $y \sim \text{BEINF}(\pi,\gamma,\mu,\phi)$, uma variável aleatória que assume valores no intervalo [0,1] e possui distribuição beta inflacionada em zero e um, com fdp na forma apresentada em (2.2.6).

A média e a variância da distribuição são dadas por

$$\text{E}(y) = \pi\gamma + (1-\pi)\mu,$$

$$\text{Var}(y) = \pi\gamma(1-\gamma) + (1-\pi)\frac{V(\mu)}{\phi+1} + \pi(1-\pi)(\gamma-\mu)^2,$$

em que $V(\mu) = \mu(1-\mu)$. De forma análoga, observa-se que permanece válida a interpretação de ϕ como um parâmetro de precisão.

A Figura 5 apresenta curvas para a função densidade da distribuição BEINF considerando alguns valores de μ, ϕ, e com π e γ fixados. Conforme pode se observar, as curvas assumem formas simétricas quando $\mu = 0{,}5$ e $\gamma = 0{,}5$ ou assimétricas, quando $\mu \neq 0{,}5$. Para $\mu = 0{,}5$ e $\phi < 2$ a parte da curva compreendida no intervalo (0,1) assume forma de 'U'. Além disso, vemos que as curvas apresentam massa de probabilidade em ambos os pontos extremos. Também observa-se que para $\phi \leq 2$, as curvas apresentam forma de 'J' ou 'J' invertido quando $\mu > 0{,}5$ ou $\mu < 0{,}5$, respectivamente.

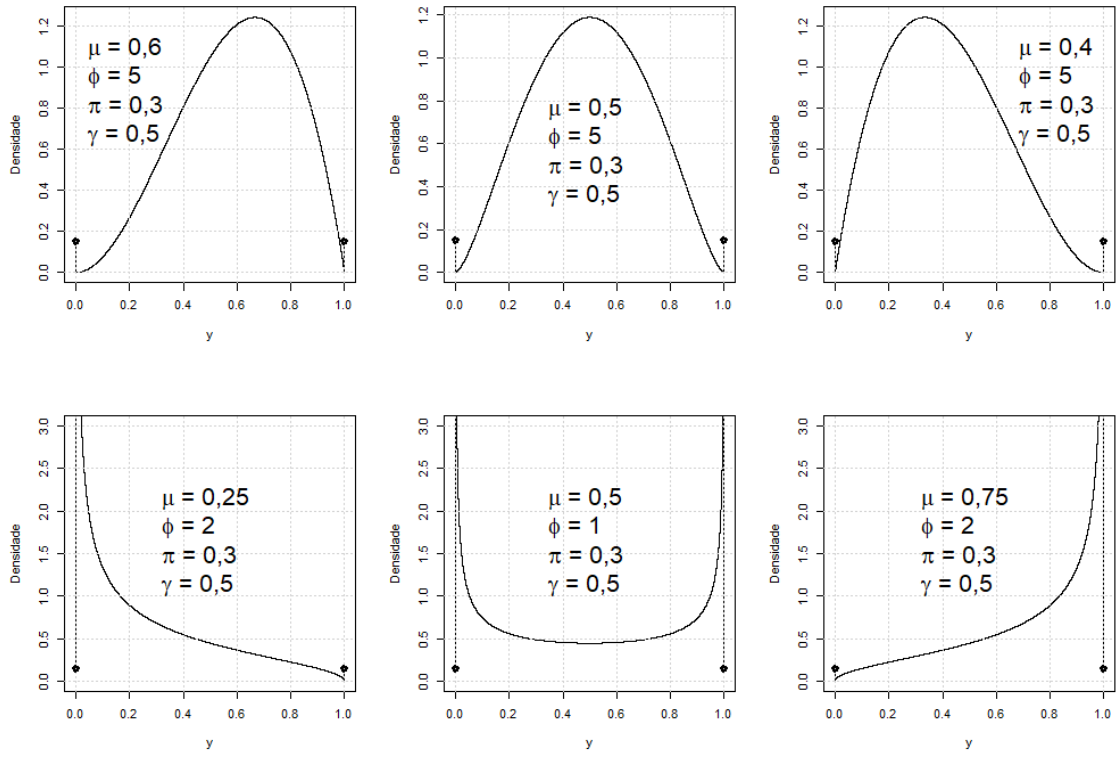

Figura 5 Curvas para a fdp da distribuição BEINF para diferentes valores de ϕ, μ, e π e γ fixados.

Ospina e Ferrari (2010) apresentam outra parametrização para a função densidade de probabilidade em (2.2.6). Fazendo $\delta_1 = \pi\gamma$ e $\delta_0 = \pi - \delta_1$, obtemos

$$\text{beinf}^*(y;\delta_0,\delta_1,\mu,\phi) = \begin{cases} \delta_0, & \text{se } \ y = 0, \\ \delta_1, & \text{se } \ y = 1, \\ (1-\delta_0-\delta_1)f(y;\mu,\phi), & \text{se } \ 0 < y < 1. \end{cases} \tag{2.2.8}$$

Note que a parametrização em (2.2.8) exige que a condição $0 < \delta_0 + \delta_1 < 1$ seja satisfeita. Essa segunda forma da função densidade da BEINF permite uma interpretação mais direta, e é particularmente útil quando considerada em um contexto de regressão, uma vez que os parâmetros δ_0 e δ_1 representam, respectivamente, a probabilidade da variável aleatória y assumir o valor 0, $\delta_0 = P(y = 0)$ ou o valor 1, $\delta_1 = P(y = 1)$.

3 Modelos de regressão beta

3.1 Regressão beta com precisão constante

A reparametrização da distribuição beta introduzida por Ferrari e Cribari-Neto (2004) viabilizou a sua utilização em modelos de regressão. Dada a expressão (2.1.4), segundo Ferrari e Cribari-Neto (2004) o modelo de regressão beta é obtido assumindo que, para n realizações independentes de uma variável aleatória y com distribuição beta, a média μ_t de cada observação y_t pode ser escrita como

$$g_\mu(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = X_t^\top \beta = \eta_t, \tag{3.1.1}$$

em que $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^\top \in \mathbb{R}^k$ é um vetor de parâmetros desconhecidos associado à média, $X_t = (x_{t1}, x_{t2}, \ldots, x_{tk})^\top \in \mathbb{R}^k$ é o vetor de valores conhecidos das k variáveis explicativas (covariáveis) para a t-ésima observação ($t = 1,2,\ldots,n$), e $g_\mu(.)$ é uma função de ligação contínua, estritamente monótona e duas vezes diferenciável. O principal objetivo associado a $g_\mu(.)$ é restringir μ_t ao suporte da distribuição beta que é o intervalo (0,1). O requisito de que a função de ligação $g_\mu(.)$ seja duas vezes diferenciável viabiliza o processo de estimação, em particular, a obtenção da matriz de informação de Fisher. Tal matriz é necessária para dimensionar a variabilidade assintótica das estimativas dos parâmetros de regressão, conforme será visto mais adiante.

Para o modelo de regressão em estudo existem diversas opções para a função de ligação $g_\mu(.)$ que atendem aos requisitos mencionados. A rigor, a inversa da função de distribuição acumulada de qualquer distribuição contínua poderia ser utilizada, entretanto, as funções de ligação mais citadas e utilizadas (FERRARI; CRIBARI-NETO, 2004; OSPINA, 2004; PEREIRA, 2010) estão a seguir:

- função logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, cuja inversa corresponde à fda da distribuição logística padrão.
- função probit: $g(\mu) = \Phi^{-1}(\mu)$, em que $\Phi(.)$ é a fda da distribuição normal padrão.
- função log-log: $g(\mu) = -\log[-\log(\mu)]$, em que $g^{-1}(\mu)$ é a fda da distribuição Gumbel padrão (máximo), correspondente a uma das duas formas da distribuição do valor extremo padrão, tipo I (GUMBEL, 1954).
- função complementar log-log: $g(\mu) = \log[-\log(1-\mu)]$, em que $g^{-1}(\mu)$ é a fda da distribuição Gumbel padrão (mínimo), correspondente a uma segunda forma da distribuição do valor extremo padrão, tipo I (GUMBEL, 1954).

- função Cauchit: $g(\mu) = \tan[\pi(\mu - 0.5)]$, cuja inversa corresponde à fda da distribuição Cauchy padrão.

O vetor de parâmetros β da regressão e o parâmetro de precisão ϕ são desconhecidos e, portanto, devem ser estimados. Para este caso, Ferrari e Cribari-Neto (2004) utilizaram o método da máxima verossimilhança, por meio do qual são estimados os valores dos parâmetros que maximizam a função densidade de probabilidade conjunta da amostra. Conforme será visto mais adiante, o estimador de máxima verossimilhança possui propriedades assintóticas úteis para o processo inferencial que será feito.

Sejam $y_1, y_2, \ldots, y_n$ variáveis aleatórias independentes tal que $y_t \sim \mathcal{B}(\mu_t,\phi)$. Então, a função de verossimilhança para $\theta = (\beta^\top, \phi)^\top$ é dada por

$$L(\theta) = \prod_{t=1}^{n} f(y_t; \mu_t,\phi) = \prod_{t=1}^{n} \left[\frac{\Gamma(\mu_t)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y_t^{\mu_t\phi-1}(1-y_t)^{(1-\mu_t)\phi-1} \right].$$

Assim, o logaritmo da função de verossimilhança para θ é

$$\begin{aligned} l(\theta) &= \log(L(\theta)) \\ &= \sum_{t=1}^{n} \log(f(y_t; \mu_t,\phi)) \\ &= \sum_{t=1}^{n} \log \left[\frac{\Gamma(\mu_t)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y_t^{\mu_t\phi-1}(1-y_t)^{(1-\mu_t)\phi-1} \right] \\ &= \sum_{t=1}^{n} l_t(\mu_t,\phi). \end{aligned} \tag{3.1.2}$$

Desenvolvendo a expressão dentro do somatório, e utilizando as propriedades do logaritmo natural, obtemos que

$$\begin{aligned} l_t(\mu_t,\phi) = {} & \log\Gamma(\phi) - \log\Gamma(\mu_t\phi) - \log\Gamma((1-\mu_t)\phi) \\ & + (\mu_t\phi - 1)\log(y_t) + [(1-\mu_t)\phi - 1]\log(1-y_t), \end{aligned}$$

com $\mu_t = g^{-1}(\eta_t)$ uma função que depende de β e X_t.

O método de máxima verossimilhança consiste em obter o valor de $\theta = (\beta^\top, \phi)^\top$ que maximiza a expressão em (3.1.2) e, para isso, é necessário obter as derivadas parciais de $l_t(\mu_t,\phi)$ com relação a cada um dos parâmetros β e ϕ. Ressalta-se que, nesse caso, a média μ_t é estimada indiretamente através de β por meio da estrutura de regressão associada a μ_t, conforme definido em (3.1.1). O parâmetro de precisão ϕ é estimado diretamente e, portanto, mantido constante para todas as observações. As derivadas parciais referentes aos parâmetros β e ϕ mencionadas no parágrafo anterior são chamadas de vetores escore para β e ϕ, aqui representadas por $U_\beta(\theta)$ e $U_\phi(\theta)$, respectivamente.

Assim, para $i = 1, \ldots, k$, as entradas do vetor escore para β são calculadas a partir de

$$\begin{aligned} U_{\beta_i}(\theta) &= \frac{\partial l(\theta)}{\partial \beta_i} \\ &= \sum_{t=1}^{n} \frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t} \frac{d\mu_t}{d\eta_t} \frac{\partial \eta_t}{\partial \beta_i} \\ &= \sum_{t=1}^{n} \left\{ \phi \left[\log(y_t) - \log(1-y_t) - \psi(\mu_t\phi) + \psi((1-\mu_t)\phi) \right] \frac{1}{{g_\mu}'(\mu_t)} x_{ti} \right\} \\ &= \sum_{t=1}^{n} \phi({y_t}^* - {\mu_t}^*) \frac{1}{{g_\mu}'(\mu_t)} x_{ti}, \end{aligned}$$

em que ${g_\mu}'(\mu_t)$ é a primeira derivada de $g_\mu(.)$ avaliada em μ_t, $\psi(\lambda)$ denota a função digama, isto é, $\psi(\lambda) = \partial \log \Gamma(\lambda)/\partial\lambda$, ${y_t}^* = \log(y_t/(1-y_t))$ e ${\mu_t}^* = \mathrm{E}(y_t^*) = \psi(\mu_t\phi) - \psi((1-\mu_t)\phi)$.

Seja X uma matriz de dimensão $n \times k$ em que cada coluna de X representa os valores conhecidos da i-ésima covariável, $i = 1, 2, \ldots, k$, $y^* = ({y_1}^*, \ldots, {y_n}^*)^\top$, $\mu^* = ({\mu_1}^*, \ldots, {\mu_n}^*)^\top$ e $T = \mathrm{diag}\{1/{g_\mu}'(\mu_1), \ldots, 1/{g_\mu}'(\mu_n)\}$. Então, o vetor escore $U_\beta(\theta)$ pode ser representado por

$$U_\beta(\theta) = \phi X^\top T(y^* - \mu^*). \tag{3.1.3}$$

O vetor escore para ϕ contém uma única entrada, que é expressa como

$$\begin{aligned} U_\phi(\theta) &= \frac{\partial l(\theta)}{\partial \phi} \\ &= \sum_{t=1}^{n} \Big\{ \mu_t[\log(y_t) - \log(1-y_t) - \psi(\mu_t\phi) \\ &+ \psi((1-\mu_t)\phi)] + \log(1-y_t) + \psi(\phi) - \psi[(1-\mu_t)\phi] \Big\} \\ &= \sum_{t=1}^{n} \Big\{ \mu_t[{y_t}^* - {\mu_t}^*] + \log(1-y_t) + \psi(\phi) - \psi[(1-\mu_t)\phi] \Big\}. \end{aligned}$$

Por fim, o estimador de máxima verossimilhança de θ é obtido por meio da resolução do sistema de equações

$$U_\beta(\theta) = 0,$$

$$U_\phi(\theta) = 0,$$

com relação a θ.

O estimador de máxima verossimilhança para $\theta = (\beta^\top, \phi)^\top$ será denotado por $\hat{\theta} = (\hat{\beta}^\top, \hat{\phi})^\top$. Observa-se que não é possível explicitar tais estimadores de forma analítica, uma vez que não possuem forma fechada. Sendo assim, é necessário recorrer a métodos de otimização não linear para maximizar o logaritmo da função de verossimilhança e,

portanto, obter numericamente as estimativas para θ. Ferrari e Cribari-Neto (2004) citam como exemplo os algoritmos de *Newton-Raphson* ou um algoritmo *Quasi-Newton*, e, inclusive, sugerem valores iniciais para o processo de convergência.

Obtidas as estimativas pontuais para os parâmetros do modelo, se torna necessário determinar o comportamento da variabilidade desses estimadores e, desse modo, construirmos estimativas intervalares e testarmos hipóteses sobre os parâmetros. Para isso, utilizamos a propriedade de normalidade assintótica dos estimadores de máxima verossimilhança (BICKEL; DOKSUM, 2001). Sob condições usuais de regularidade, e para amostras grandes, pode-se demonstrar que

$$\widehat{\theta} = \begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \overset{a}{\sim} N_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}; K^{-1} \right),$$

em que $\overset{a}{\sim}$ denota que a distribuição é aproximada e K é a Matriz de Informação de Fisher dada por

$$K = K(\theta) = \begin{bmatrix} -E\left(\frac{\partial^2 l(\theta)}{\partial\beta\partial\beta^\top}\right) & -E\left(\frac{\partial^2 l(\theta)}{\partial\beta\partial\phi}\right) \\ -E\left(\frac{\partial^2 l(\theta)}{\partial\phi\partial\beta^\top}\right) & -E\left(\frac{\partial^2 l(\theta)}{\partial\phi^2}\right) \end{bmatrix} = \begin{bmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{bmatrix},$$

em que $K_{\beta\beta} = \phi X^\top W X$, $K_{\beta\phi} = K_{\phi\beta}^\top = X^\top T c$, $K_{\phi\phi} = \mathrm{tr}(D)$ e

$$\begin{aligned} W &= \mathrm{diag}\{w_1, \ldots, w_n\}, \\ D &= \mathrm{diag}\{d_1, \ldots, d_n\}, \\ c &= (c_1, \ldots, c_n)^\top, \\ w_t &= \phi\{\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)\}\left\{\frac{1}{g_\mu{}'(\mu_t)}\right\}^2, \\ d_t &= \psi'(\mu_t\phi){\mu_t}^2 + \psi'((1-\mu_t)\phi)(1-\mu_t)^2 - \psi'(\phi), \\ c_t &= \phi\{\psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)\}. \end{aligned}$$

Segundo Ferrari e Cribari-Neto (2004), a inversa da Matriz de Informação de Fisher é da forma

$$\begin{aligned} K^{-1} &= K^{-1}(\theta) = \begin{bmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{bmatrix}, \quad \text{com} \\ K^{\beta\beta} &= \frac{1}{\phi}(X^\top W X)^{-1}\left\{I_k + \frac{X^\top T c c^\top T^\top X (X^\top W X)^{-1}}{\varrho\phi}\right\}, \\ K^{\beta\phi} &= (K^{\phi\beta})^\top = -\frac{1}{\varrho\phi}(X^\top W X)^{-1} X^\top T c\ , \\ K^{\phi\phi} &= \varrho^{-1}, \end{aligned}$$

em que $\varrho = \text{tr}(D) - \phi^{-1}c^\top T^\top X(X^\top WX)^{-1}X^\top Tc$, I_k é uma matriz identidade de dimensão k.

3.2 Regressão beta com precisão variável

Na seção anterior foi visto que, apesar de heterocedástico, o modelo proposto por Ferrari e Cribari-Neto (2004) considera que a precisão é constante para todas as observações, o que nem sempre será apropriado supor. Conforme discorrido por Bayer (2011), a utilização de modelos com precisão constante quando esta for variável, pode levar a uma estimação inadequada do parâmetro ϕ.

No contexto dos modelos lineares generalizados (MLGs) introduzidos por Nelder e Wedderburn (1972), existem trabalhos onde são considerados os modelos lineares generalizados duplos, nos quais a média e a precisão são modeladas simultaneamente (NELDER; LEE, 1991; SMYTH; VERBYLA, 1999). Nesse sentido, Smithson e Verkuilen (2006) propuseram uma extensão ao modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004). Sob essa nova abordagem, adicionou-se uma estrutura de regressão linear para modelar também o parâmetro de precisão ϕ.

Consideraremos que a precisão ϕ_t será modelada simultaneamente à média μ_t, por meio da estrutura de regressão

$$g_\phi(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = Z_t^\top \gamma = \vartheta_t, \tag{3.2.1}$$

em que $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^\top \in \mathbb{R}^q$ é um vetor de parâmetros desconhecidos associado à precisão, $Z_t = (z_{t1}, z_{t2}, \ldots, z_{tq})^\top \in \mathbb{R}^q$ é o vetor de valores conhecidos das q covariáveis da precisão para a t-ésima observação, e $g_\phi(.)$ é uma função de ligação contínua, estritamente monótona e duas vezes diferenciável. Observa-se que, diferente do que ocorre com a média, que deve ser mapeada no domínio da variável resposta, o parâmetro de precisão deve assumir valores estritamente positivos, uma vez que $\text{Var}(y)$ não pode ser negativa. Dentre as funções de ligação que atendem a esses critérios, são citadas na literatura (SMITHSON; VERKUILEN, 2006) as funções de ligação abaixo:

- função logaritmo: $g_\phi(\phi) = \log(\phi)$.
- função raiz-quadrada: $g_\phi(\phi) = \sqrt{\phi}$.

A função de ligação $g_\phi(\cdot)$ mais utilizada é a logarítmica, por meio da qual obtemos que

$$\phi_t = \exp\left[\sum_{i=1}^{q} z_{tj}\gamma_j\right] = e^{\vartheta_t}.$$

O processo de estimação por máxima verossimilhança é semelhante ao adotado para o modelo com precisão constante mas, nesse caso, a estimação da precisão também é efetuada em função dos parâmetros integrantes da estrutura de regressão, conforme (3.1.1) e (3.2.1).

Tomando $y_1, y_2, \ldots, y_n$ variáveis aleatórias independentes tal que $y_t \sim \mathcal{B}(\mu_t,\phi_t)$, a função de verossimilhança para $\theta = (\beta^\top, \gamma^\top)^\top$ é dada por

$$\begin{aligned} L(\theta) &= \prod_{t=1}^{n} f(y_t; \mu_t,\phi_t) \\ &= \prod_{t=1}^{n} \left[\frac{\Gamma(\mu_t)}{\Gamma(\mu_t\phi_t)\Gamma((1-\mu_t)\phi_t)} y_t^{\mu_t\phi_t - 1}(1-y_t)^{(1-\mu_t)\phi_t - 1} \right], \end{aligned}$$

e o respectivo logaritmo da função de verossimilhança para θ é

$$\begin{aligned} l(\theta) &= \log(L(\theta)) \\ &= \sum_{t=1}^{n} \log(f(y_t; \mu_t,\phi_t)) \\ &= \sum_{t=1}^{n} \log \left[\frac{\Gamma(\mu_t)}{\Gamma(\mu_t\phi_t)\Gamma((1-\mu_t)\phi_t)} y_t^{\mu_t\phi_t - 1}(1-y_t)^{(1-\mu_t)\phi_t - 1} \right] \\ &= \sum_{t=1}^{n} l_t(\mu_t,\phi_t), \end{aligned} \tag{3.2.2}$$

sendo

$$\begin{aligned} l_t(\mu_t,\phi_t) = {} & \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) \\ & + (\mu_t\phi_t - 1)\log(y_t) + [(1-\mu_t)\phi_t - 1]\log(1-y_t), \end{aligned} \tag{3.2.3}$$

com $\mu_t = g_\mu^{-1}(\eta_t)$, uma função de β e X_t, e $\phi_t = g_\phi^{-1}(\vartheta_t)$, uma função de γ e Z_t.

As entradas do vetor escore para β, $U_{\beta_i}(\theta)$, $i = 1,2,\ldots,k$, são dadas por

$$\begin{aligned} U_{\beta_i}(\theta) &= \frac{\partial l(\theta)}{\partial \beta_i} \\ &= \sum_{t=1}^{n} \frac{\partial l_t(\mu_t,\phi_t)}{\partial \mu_t} \frac{d\mu_t}{d\eta_t} \frac{\partial \eta_t}{\partial \beta_i} \\ &= \sum_{t=1}^{n} \left\{ \phi_t \left[\log(y_t) - \log(1-y_t) - \psi(\mu_t\phi_t) + \psi((1-\mu_t)\phi_t) \right] \frac{1}{g_\mu{}'(\mu_t)} x_{ti} \right\} \\ &= \sum_{t=1}^{n} \phi_t (y_t^* - \mu_t^*) \frac{1}{g_\mu{}'(\mu_t)} x_{ti}. \end{aligned}$$

Definindo a matriz $\Phi = \text{diag}\{\phi_1, \ldots, \phi_n\}$, obtemos que o vetor escore para β é

$$U_\beta(\theta) = \Phi X^\top T(y^* - \mu^*).$$

As entradas do vetor escore para γ, $U_{\gamma_j}(\theta)$, com $j = 1,2,\ldots,q$, são dadas por

$$\begin{aligned}
U_{\gamma_j}(\theta) &= \frac{\partial l(\theta)}{\partial \gamma_j} \\
&= \sum_{t=1}^{n} \frac{\partial l_t(\mu_t,\phi_t)}{\partial \phi_t} \frac{d\phi_t}{d\vartheta_t} \frac{\partial \vartheta_t}{\partial \gamma_j} \\
&= \sum_{t=1}^{n} \Big\{ \mu_t[\log(y_t) - \log(1-y_t) - \psi(\mu_t\phi_t) \\
&+ \psi((1-\mu_t)\phi_t)] + \log(1-y_t) + \psi(\phi_t) - \psi[(1-\mu_t)\phi_t] \frac{1}{g_\phi{}'(\phi_t)} z_{tj} \Big\} \\
&= \sum_{t=1}^{n} \left\{ \mu_t(y_t{}^* - \mu_t{}^*) + \log(1-y_t) + \psi(\phi_t) - \psi((1-\mu_t)\phi_t) \frac{1}{g_\phi{}'(\phi_t)} z_{tj} \right\} \\
&= \sum_{t=1}^{n} a_t \frac{1}{g_\phi{}'(\phi_t)} z_{tj},
\end{aligned}$$

em que $a_t = \mu_t(y_t{}^* - \mu_t{}^*) + \log(1-y_t) + \psi(\phi_t) - \psi((1-\mu_t)\phi_t)$.

Seja Z uma matriz de dimensão $n \times q$ em que cada coluna de Z representa os n valores observados para a j-ésima covariável, $j = 1,2,\ldots,q$, $H = \text{diag}\{1/g_\phi{}'(\phi_1), \ldots, 1/g_\phi{}'(\phi_n)\}$ e $a = (a_1, \ldots, a_n)^\top$, temos que o vetor escore para γ é

$$U_\gamma(\theta) = HZ^\top a. \tag{3.2.4}$$

O estimador de máxima verossimilhança para θ, denotado por $\hat{\theta} = (\hat{\beta}^\top, \hat{\gamma}^\top)^\top$, é obtido resolvendo o sistema de equações

$$U_\beta(\theta) = 0,$$

$$U_\gamma(\theta) = 0,$$

com relação a $\theta = (\beta^\top, \gamma^\top)^\top$. Assim como ocorre na Subseção 3.1, nesta situação também não será possivel explicitar os estimadores de máxima verossimilhança dos parâmetros de regressão β e γ, denotados por $\hat{\beta}$ e $\hat{\gamma}$ de forma analítica, sendo necessário recorrer a métodos iterativos de estimação.

Sob condições de regularidade e para n suficientemente grande, Espinheira (2007) define que a distribuição conjunta aproximada de $\hat{\theta} = (\hat{\beta}^\top, \hat{\gamma}^\top)^\top$ é normal $(k+q)$-variada,

isto é

$$\hat{\theta} = \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \overset{a}{\sim} N_{k+q}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}; K^{*-1} \right),$$

em que K^* é a Matriz de Informação de Fisher dos parâmetros, representada por

$$K^* = K^*(\theta) = \begin{bmatrix} -E\left(\frac{\partial^2 l(\theta)}{\partial\beta\partial\beta^\top}\right) & -E\left(\frac{\partial^2 l(\theta)}{\partial\beta\partial\gamma^\top}\right) \\ -E\left(\frac{\partial^2 l(\theta)}{\partial\gamma\partial\beta^\top}\right) & -E\left(\frac{\partial^2 l(\theta)}{\partial\gamma\partial\gamma^\top}\right) \end{bmatrix} = \begin{bmatrix} K^*_{\beta\beta} & K^*_{\beta\gamma} \\ K^*_{\gamma\beta} & K^*_{\gamma\gamma} \end{bmatrix},$$

sendo $K^*_{\beta\beta} = X^\top \Phi W X$, $K^*_{\beta\gamma} = K^{*\top}_{\gamma\beta} = X^\top C T H Z$, $K^*_{\gamma\gamma} = Z^\top D^* Z$ e

$$\begin{aligned} \Phi &= \text{diag}\{\phi_1, \ldots, \phi_n\}, \\ C &= \text{diag}\{{c_1}^*, \ldots, {c_n}^*\}, \\ D^* &= \text{diag}\{{d_1}^*, \ldots, {d_n}^*\}, \\ {c_t}^* &= \phi_t\{\psi'(\mu_t\phi_t)\mu_t - \psi'((1-\mu_t)\phi_t)(1-\mu_t)\}, \\ {d_t}^* &= [\psi'(\mu_t\phi_t){\mu_t}^2 + \psi'((1-\mu_t)\phi_t)(1-\mu_t)^2 - \psi'(\phi_t)][g_\phi(\phi_t)]^{-2}. \end{aligned}$$

A inversa da Matriz de Informação de Fisher é da forma (ESPINHEIRA, 2007)

$$K^{*-1} = K^{*-1}(\theta) = \begin{bmatrix} K^{*\beta\beta} & K^{*\beta\gamma} \\ K^{*\gamma\beta} & K^{*\gamma\gamma} \end{bmatrix} \tag{3.2.5}$$

em que

$$\begin{aligned} K^{*\beta\beta} &= \left(X^\top \Phi W^* X - X^\top C T H Z (Z^\top D^* Z)^{-1} Z^\top H T C^\top X\right)^{-1}, \\ K^{*\beta\gamma} &= (K^{*\gamma\beta})^\top = -K^{*\beta\beta} X^\top C T H Z (Z^\top D^* Z)^{-1}, \\ K^{*\gamma\gamma} &= (Z^\top D^* Z)^{-1}\left\{I_q + (Z^\top H T C^\top X) K^{*\beta\beta} X^\top C T H Z (Z^\top D^* Z)^{-1}\right\}, \end{aligned}$$

sendo $W^* = \text{diag}\{{w_1}^*, \ldots, {w_n}^*\}$, ${w_t}^* = \phi_t\{\psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t)\}\{d\mu_t/d\eta_t\}^2$ e I_q uma matriz identidade de ordem q, X e T conforme definidas na Seção (3.1).

Espinheira (2007) demonstra que quando $\phi_1 = \phi_2 = \ldots = \phi_n = \phi$, ou seja, no caso em que a precisão é tratada como constante ao longo das observações da amostra, a expressão em (3.2.5) se reduz à Matriz de Informação de Fisher em (3.1), referente ao modelo de regressão beta com precisão constante.

3.3 Regressão beta inflacionada

As distribuições beta inflacionadas introduzidas nas Seções 2.2.1 e 2.2.2 viabilizaram o desenvolvimento de modelos de regressão baseados na distribuição beta e que

permitem modelar, além de dados limitados ao intervalo contínuo unitário, também aqueles que contém valores em zero e/ou em um.

3.3.1 Regressão beta inflacionada em zero ou em um

Conforme Ospina e Ferrari (2012), assumindo $y_1, y_2, \ldots, y_n$ variáveis aleatórias independentes com função densidade de probabilidade conforme (2.2.1), e com parâmetros $\alpha = \alpha_t$, $\mu = \mu_t$ e $\phi = \phi_t$, as estruturas de regressão para os parâmetros são dadas por $g_\mu(\mu_t)$, definida conforme (3.1.1), $g_\phi(\phi_t)$, definida conforme (3.2.1), e

$$g_\alpha(\alpha_t) = \sum_{i=1}^{p} v_{ti}\rho_i = V_t^{\top}\rho = \lambda_t, \tag{3.3.1}$$

em que $\rho = (\rho_1, \rho_2, \ldots, \rho_p)^{\top} \in \mathbb{R}^p$ é um vetor de parâmetros desconhecido associado a α_t, $v_{ti} = (v_{t1}, v_{t2}, \ldots, v_{tp})^{\top} \in \mathbb{R}^p$ é o vetor de valores conhecidos das p covariáveis de α para a t-ésima observação, e $g_\alpha(.)$ é uma função de ligação contínua, estritamente monótona e duas vezes diferenciável tal que $g_\alpha : (0,1) \to \mathbb{R}$. As mesmas opções de funções de ligação enumeradas na Subseção 3.1, e aplicáveis a $g_\mu(.)$ podem ser utilizadas também para $g_\alpha(\cdot)$.

Sob a especificação definida no parágrafo anterior, o modelo inflacionado no ponto $c = 0$ é chamado de regressão beta inflacionada em zero (regressão BEZI) e quando a inflação ocorre em $c = 1$ é chamado de regressão beta inflacionada em um (regressão BEOI).

A definição do modelo na forma estabelecida em (3.3.1) é bastante útil pois proporciona a interpretação dos parâmetros de regressão associados a α_t de forma direta.

A função de verossimilhança para $\theta = (\rho^{\top}, \beta^{\top}, \gamma^{\top})^{\top}$ é

$$\begin{aligned} L(\theta) &= \prod_{t=1}^{n} bi_c(y_t; \alpha_t, \mu_t, \phi_t) \\ &= \prod_{t=1}^{n} \left[\left\{ \alpha_t^{I_{\{c\}}(y_t)} (1-\alpha_t)^{1-I_{\{c\}}(y_t)} \right\} \left\{ f(y_t; \mu_t, \phi_t)^{1-I_{\{c\}}(y_t)} \right\} \right] \\ &= \left[\prod_{t=1}^{n} \left\{ \alpha_t^{I_{\{c\}}(y_t)} (1-\alpha_t)^{1-I_{\{c\}}(y_t)} \right\} \right] \left[\prod_{t:y_t \in (0,1)} f(y_t; \mu_t, \phi_t) \right]. \end{aligned}$$

Seja $\theta = (\theta_1^{\top}, \theta_2^{\top})^{\top}$, $\theta_1 = \rho$ e $\theta_2 = (\beta^{\top}, \gamma^{\top})^{\top}$, então o logaritmo da função de verossimi-

lhança para θ é dado por

$$\begin{aligned} l(\theta) &= \log(L(\theta)) \\ &= l_1(\theta_1) + l_2(\theta_2) \\ &= \sum_{t=1}^{n} l_t(\alpha_t) + \sum_{t:y_t\in(0,1)} l_t(\mu_t,\phi_t), \end{aligned} \tag{3.3.2}$$

em que

$$\begin{aligned} l_t(\alpha_t) &= I_{\{c\}}(y_t)\log(\alpha_t) + (1 - I_{\{c\}}(y_t))\log(1-\alpha_t), \\ l_t(\mu_t,\phi_t) &= \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) + (\mu_t\phi_t - 1)\log(y_t) \\ &+ [(1-\mu_t)\phi_t - 1]\log(1-y_t) \\ &= \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) \\ &+ (\mu_t\phi_t - 1)[\log(y_t) - \log(1-y_t)] + \phi_t\log(1-y_t) - 2\log(1-y_t) \\ &= \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) \\ &+ (\mu_t\phi_t - 1)\log\left(\frac{y_t}{1-y_t}\right) + (\phi_t - 2)\log(1-y_t). \end{aligned}$$

Conforme ressaltado por Ospina e Ferrari (2012), temos que $\sum_{t=1}^{n} l_t(\alpha_t)$ representa o logaritmo da função de verossimilhança de um modelo de regressão linear para respostas binárias, cuja probabilidade de sucesso para a t-ésima observação é $\alpha_t = {g_\alpha}^{-1}(\lambda_t)$. A quantidade $\sum_{t:y_t\in(0,1)} l_t(\mu_t,\phi_t)$ é o logaritmo da função de verossimilhança de um modelo de regressão beta baseado nas observações restritas ao intervalo contínuo (0,1).

Observe que o modelo especificado em (3.3.1) considera a precisão variável ao longo da amostra, e por esse motivo é estabelecida uma estrutura de regressão para modelar o parâmetro ϕ. Desse modo, a forma obtida para $l_t(\mu_t,\phi_t)$ é a mesma verificada em (3.2.3), referente ao modelo de regressão com precisão variável (SMITHSON; VERKUILEN, 2006). Caso estivesse sendo considerada a precisão constante, então não seria estabelecida uma estrutura de regressão pra modelar ϕ, e portanto o logaritmo da função de verossimilhança para o componente seria da forma (3.1.2).

Os vetores escore $U_\rho(\theta)$, $U_\beta(\theta)$ e $U_\gamma(\theta)$ são obtidos diferenciando a expressão em (3.3.2) em relação a cada um dos vetores de parâmetros de regressão desconhecidos ρ, β e γ, respectivamente. Assim, considerando $l = 1,\ldots,p$, $i = 1,\ldots,k$ e $j = 1,\ldots,q$,

obtém-se que

$$\begin{aligned}
U_{\rho_l}(\theta) &= \frac{\partial l_1(\theta_1)}{\partial \rho_l} \\
&= \sum_{t=1}^{n} \frac{\partial l_t(\alpha_t)}{\partial \alpha_t}\frac{d\alpha_t}{d\lambda_t}\frac{\partial \lambda_t}{\partial \rho_l} \\
&= \sum_{t=1}^{n} \frac{I_{(c)}(y_t)-\alpha_t}{\alpha_t(1-\alpha_t)}\frac{1}{{g_\alpha}'(\alpha_t)}v_{tl}, \\
U_{\beta_i}(\theta) &= \frac{\partial l_2(\theta_2)}{\partial \beta_i} \\
&= \sum_{t:y_t\in(0,1)} \frac{\partial l_t(\mu_t,\phi_t)}{\partial \mu_t}\frac{d\mu_t}{d\eta_t}\frac{\partial \eta_t}{\partial \beta_i} \\
&= \sum_{t:y_t\in(0,1)} \left\{\phi_t\left[\log(y_t)-\log(1-y_t)-\psi(\mu_t\phi_t)+\psi((1-\mu_t)\phi_t)\right]\frac{1}{{g_\mu}'(\mu_t)}x_{ti}\right\} \\
&= \sum_{t:y_t\in(0,1)} \phi_t({y_t}^*-{\mu_t}^*)\frac{1}{{g_\mu}'(\mu_t)}x_{ti} \\
&= \sum_{t=1}^{n} \left[1-I_{\{c\}}(y_t)\right]\phi_t({y_t}^*-{\mu_t}^*)\frac{1}{{g_\mu}'(\mu_t)}x_{ti}, \\
U_{\gamma_j}(\theta) &= \frac{\partial l_2(\theta_2)}{\partial \gamma_j} \\
&= \sum_{t:y_t\in(0,1)} \frac{\partial l_t(\mu_t,\phi_t)}{\partial \phi_t}\frac{d\phi_t}{d\vartheta_t}\frac{\partial \vartheta_t}{\partial \gamma_j} \\
&= \sum_{t:y_t\in(0,1)} \Big\{\mu_t\left[\log(y_t)-\log(1-y_t)-\psi(\mu_t\phi_t)+\psi((1-\mu_t)\phi_t)\right] \\
&+ \log(1-y_t)+\psi(\phi_t)-\psi((1-\mu_t)\phi_t)\frac{1}{{g_\phi}'(\phi_t)}z_{tj}\Big\} \\
&= \sum_{t:y_t\in(0,1)} \left\{\mu_t({y_t}^*-{\mu_t}^*)+s(y_t)+\psi(\phi_t)-\psi((1-\mu_t)\phi_t)\frac{1}{{g_\phi}'(\phi_t)}z_{tj}\right\} \\
&= \sum_{t:y_t\in(0,1)} a_t\frac{1}{{g_\phi}'(\phi_t)}z_{tj} \\
&= \sum_{t=1}^{n} \left[1-I_{\{c\})}(y_t)\right]a_t\frac{1}{{g_\phi}'(\phi_t)}z_{tj},
\end{aligned}$$

em que ${\mu_t}^* = \mathrm{E}({y_t}^*|I_{(0,1)}(y_t)=1) = \psi(\mu_t\phi_t)-\psi((1-\mu_t)\phi_t)$,

$${y_t}^* = \begin{cases} \log\left(\frac{y_t}{1-y_t}\right), & \text{se } \ y_t\in(0,1), \\ 0, & \text{caso contrário,} \end{cases}$$

$$s(y_t) = \begin{cases} \log(1-y_t), & \text{se } \ y_t \in (0,1), \\ 0, & \text{caso contrário.} \end{cases}$$

Adicionalmente às definições efetuadas na Subseção 3.2, considere V uma matriz de dimensão $n \times p$ em que cada coluna de V representa os n valores observados para a j-ésima covariável, $j = 1,2,\ldots,p$, $y^c = (I_{\{c\}}(y_1),\ldots,I_{\{c\}}(y_n))^\top$, $\alpha^* = (\alpha_1,\ldots,\alpha_n)^\top$, e as matrizes diagonais $M = \text{diag}\{1 - I_{\{c\}}(y_1),\ldots,1 - I_{\{c\}}(y_n)\}$, $G = \text{diag}\{1/g_\alpha{}'(\alpha_1),\ldots,$ $1/g_\alpha{}'(\alpha_n)\}$ e $P = \text{diag}\{1/[\alpha_1(1-\alpha_1))],\ldots,1/[\alpha_n(1-\alpha_n))]\}$. Os vetores escore para ρ, β e γ são dados, respectivamente, por

$$\begin{aligned} U_\rho(\theta) &= V^\top PG(y^c - \alpha^*), \\ U_\beta(\theta) &= \Phi X^\top TM(y^* - \mu^*), \\ U_\gamma(\theta) &= MHZ^\top a. \end{aligned}$$

O estimador de máxima verossimilhança para θ, denotado por $\hat\theta = (\hat\rho^\top, \hat\beta^\top, \hat\gamma^\top)^\top$, é obtido por meio da resolução do sistema de equações

$$\begin{aligned} U_\rho(\theta_1) &= 0, \\ U_\beta(\theta_2) &= 0, \\ U_\gamma(\theta_2) &= 0, \end{aligned}$$

com relação a $\theta = (\rho^\top, \beta^\top, \gamma^\top)^\top$. Não existem formas fechadas para o referido estimador de máxima verossimilhança, sendo necessário a utilização de métodos de otimização não linear como os mencionados nas Subseções 3.1 e 3.2 e, assim, obter numericamente as respectivas estimativas.

Sob condições de regularidade usuais, a distribuição conjunta aproximada de $\hat\theta = (\hat\rho^\top, \hat\beta^\top, \hat\gamma^\top)^\top$ é Normal $(p+k+q)$-variada, isto é,

$$\hat\theta = \begin{pmatrix} \hat\rho \\ \hat\beta \\ \hat\gamma \end{pmatrix} \overset{a}{\sim} N_{p+k+q}\left(\begin{pmatrix} \rho \\ \beta \\ \gamma \end{pmatrix} ; K^{**-1} \right),$$

em que K^{**} é a Matriz de Informação de Fisher dos parâmetros, que assume a forma

$$K^{**} = K^{**}(\theta) = \begin{bmatrix} K_\rho(\theta_1) & 0 \\ 0 & K_{\beta\gamma}(\theta_2) \end{bmatrix},$$

onde $K_\rho(\theta_1) = K_\rho$ é a matriz de informação de Fisher de ρ e $K_{\beta,\gamma}(\theta_2) = K_{\beta,\gamma}$ a matriz para $(\beta^\top,\gamma^\top)^\top$. A obtenção da matriz de informação de Fisher é detalhada por Ospina (2008).

3.3.2 Regressão beta inflacionada em zero e em um

Para situações em que os dados podem ser observados em ambos os extremos do intervalo unitário, ou seja, em [0,1], Ospina (2008) considera um modelo estatístico que é uma generalização natural dos modelos de regressão BEZI e BEOI.

Para simplificação dos cálculos e da notação utilizada, o modelo será especificado considerando o parâmetro de precisão ϕ como sendo constante ao longo da amostra e, portanto, sem a estrutura de regressão para modelá-lo. Assumimos $y_1, y_2, \ldots, y_n$ variáveis aleatórias independentes com fdp conforme (2.2.8) e com parâmetros $\delta_0 = \delta_{0t}$, $\delta_1 = \delta_{0t}$, $\mu = \mu_t$, e ϕ. A estrutura de regressão para μ_t é definida conforme (3.1.1) e as estruturas de regressão para δ_{0t} e δ_{1t} são dadas por

$$\begin{aligned} H(\delta_{0t},\delta_{1t}) &= (h_0(\delta_{0t},\delta_{1t}),h_1(\delta_{0t},\delta_{1t})) \\ &= \left(\sum_{i=1}^{r_0} w_{0ti}\tau_{0i}, \sum_{i=1}^{r_1} w_{1ti}\tau_{1i}\right) \\ &= ({W_{0t}}^\top \tau_0, {W_{1t}}^\top \tau_1) \\ &= (\zeta_{0t},\zeta_{1t}), \end{aligned} \tag{3.3.3}$$

em que $\delta_{0t} = P(y_t = 0)$, $\delta_{1t} = P(y_t = 1)$, $(1 - \delta_{0t} - \delta_{1t}) = P(y_t \in (0,1))$, e $h_0(\cdot,\cdot)$ e $h_1(\cdot,\cdot)$ são as funções de ligação associadas aos componentes usados para modelar δ_{0t} e δ_{1t}, respectivamente. As funções ζ_{0t} e ζ_{1t} são os preditores lineares referentes ao componente discreto da distribuição BEINF, com $\tau_0 = (\tau_{01}, \tau_{02}, \ldots, \tau_{0r_0})^\top \in \mathbb{R}^{r_0}$ e $\tau_1 = (\tau_{11}, \tau_{12}, \ldots, \tau_{1r_1})^\top \in \mathbb{R}^{r_1}$ sendo os parâmetros de regressão desconhecidos a serem estimados, e $w_{0ti} = (w_{0t1}, w_{0t2}, \ldots, w_{0tr_0})^\top \in \mathbb{R}^{r_0}$, $w_{1ti} = (w_{1t1}, w_{1t2}, \ldots, w_{1tr_1})^\top \in \mathbb{R}^{r_1}$ os vetores de valores conhecidos das r_0 e r_1 variáveis explicativas de δ_{0t} e δ_{1t}, respectivamente, para a t-ésima observação.

Ospina (2008) denomina o modelo definido em (3.3.3) de regressão beta inflacionada em zero e em um (regressão BIZU). Observa-se que os parâmetros δ_{0t} e δ_{1t} representam as probabilidades de que a variável resposta assuma o valor zero e o valor um, respectivamente. Considerando a restrição de que $0 < \delta_{0t} + \delta_{1t} < 1$, as funções de ligação $h_0(\cdot,\cdot)$ e $h_1(\cdot,\cdot)$ devem ser escolhidas de modo a satisfazer as condições $0 < \delta_{0t} < 1$ e $0 < \delta_{1t} < 1 - \delta_{0t}$, além de serem estritamente monótonas e duas vezes diferenciáveis.

A título exemplificativo, Ospina (2008) utiliza para $g_\mu(.)$ a função logit e para $h_0(\cdot,\cdot)$ e $h_1(\cdot,\cdot)$ as funções $\log(\delta_{0t}/(1-\delta_{0t}-\delta_{1t}))$ e $\log(\delta_{1t}/(1-\delta_{0t}-\delta_{1t}))$, respectivamente, chamando o modelo sob essa especificação de *modelo de regressão logístico beta inflacionado em zero e em um* (RLBIZU). Com isso, para o componente discreto do modelo

temos

$$\begin{aligned} H(\delta_{0t},\delta_{1t}) &= (h_0(\delta_{0t},\delta_{1t}),h_1(\delta_{0t},\delta_{1t})) \\ &= \left(\log\left(\frac{\delta_{0t}}{1-\delta_{0t}-\delta_{1t}}\right), \log\left(\frac{\delta_{1t}}{1-\delta_{0t}-\delta_{1t}}\right)\right), \end{aligned}$$

que implica em

$$\begin{aligned} \frac{P(y_t=0)}{P(y_t\in(0,1))} &= \frac{\delta_{0t}}{1-\delta_{0t}-\delta_{1t}} = e^{\zeta_{0t}}, \\ \frac{P(y_t=1)}{P(y_t\in(0,1))} &= \frac{\delta_{1t}}{1-\delta_{0t}-\delta_{1t}} = e^{\zeta_{1t}}, \end{aligned}$$

e, portanto

$$\begin{aligned} \delta_{0t} &= e^{\zeta_{0t}}(1-\delta_{0t}-\delta_{1t}) = \frac{e^{\zeta_{0t}}(1-\delta_{1t})}{1+e^{\zeta_{0t}}}, \\ \delta_{1t} &= e^{\zeta_{1t}}(1-\delta_{0t}-\delta_{1t}) = \frac{e^{\zeta_{1t}}(1-\delta_{0t})}{1+e^{\zeta_{1t}}}. \end{aligned}$$

Resolvendo o sistema de equações acima em relação a δ_{0t}, obtemos que

$$\begin{aligned} \delta_{0t} &= \frac{e^{\zeta_{0t}}}{1+e^{\zeta_{0t}}}\left[1-\frac{e^{\zeta_{1t}}(1-\delta_{0t})}{1+e^{\zeta_{1t}}}\right] \\ &= \frac{e^{\zeta_{0t}}}{1+e^{\zeta_{0t}}} - \frac{e^{\zeta_{0t}+\zeta_{1t}}}{(1+e^{\zeta_{0t}})(1+e^{\zeta_{1t}})} + \frac{\delta_{0t}e^{\zeta_{0t}+\zeta_{1t}}}{(1+e^{\zeta_{0t}})(1+e^{\zeta_{1t}})} \\ &= \left[\frac{e^{\zeta_{0t}}}{1+e^{\zeta_{0t}}} - \frac{e^{\zeta_{0t}+\zeta_{1t}}}{(1+e^{\zeta_{0t}})(1+e^{\zeta_{1t}})}\right]\left[\frac{(1+e^{\zeta_{0t}})(1+e^{\zeta_{1t}})}{(1+e^{\zeta_{0t}})(1+e^{\zeta_{1t}})-e^{\zeta_{0t}+\zeta_{1t}}}\right] \\ &= \left[e^{\zeta_{0t}} - \frac{e^{\zeta_{0t}+\zeta_{1t}}}{1+e^{\zeta_{1t}}}\right]\left[\frac{1+e^{\zeta_{1t}}}{(1+e^{\zeta_{0t}})(1+e^{\zeta_{1t}})-e^{\zeta_{0t}+\zeta_{1t}}}\right] \\ &= \left[\frac{e^{\zeta_{0t}}+e^{\zeta_{0t}+\zeta_{1t}}-e^{\zeta_{0t}+\zeta_{1t}}}{1+e^{\zeta_{1t}}}\right]\left[\frac{1+e^{\zeta_{1t}}}{1+e^{\zeta_{0t}}+e^{\zeta_{1t}}+e^{\zeta_{0t}+\zeta_{1t}}-e^{\zeta_{0t}+\zeta_{1t}}}\right] \\ &= \frac{e^{\zeta_{0t}}}{1+e^{\zeta_{0t}}+e^{\zeta_{1t}}}. \end{aligned}$$

De maneira análoga, resolvemos o mesmo sistema em relação a δ_{1t}, obtendo

$$\delta_{1t} = \frac{e^{\zeta_{1t}}}{1+e^{\zeta_{0t}}+e^{\zeta_{1t}}}.$$

Observe que δ_{0t} e δ_{1t} representam $P(y_t=0)$ e $P(y_t=1)$, respectivamente. Ainda,

considerando a restrição $0 < \delta_{0t} + \delta_{1t} < 1$, obtemos facilmente $P(y_t \in (0,1))$ como

$$\begin{aligned} P(y_t \in (0,1)) &= 1 - \delta_{0t} - \delta_{1t} \\ &= 1 - \frac{e^{\delta_{0t}}}{1 + e^{\delta_{0t}} + e^{\delta_{1t}}} - \frac{e^{\delta_{1t}}}{1 + e^{\delta_{0t}} + e^{\delta_{1t}}} \\ &= \frac{1}{1 + e^{\delta_{0t}} + e^{\delta_{1t}}}. \end{aligned}$$

Desse modo, ficamos com

$$\begin{aligned} \delta_{0t} = P(y_t = 0) &= \frac{e^{\delta_{0t}}}{1 + e^{\delta_{0t}} + e^{\delta_{1t}}}, \\ \delta_{1t} = P(y_t = 1) &= \frac{e^{\delta_{1t}}}{1 + e^{\delta_{0t}} + e^{\delta_{1t}}}, \\ 1 - \delta_{0t} - \delta_{1t} = P(y_t \in (0,1)) &= \frac{1}{1 + e^{\delta_{0t}} + e^{\delta_{1t}}}. \end{aligned}$$

Considerando a função densidade em (2.2.8) e a estrutura de regressão do modelo BIZU definida em (3.3.3), temos como vetor de parâmetros de regressão desconhecidos $\theta = (\theta_1^\top, \theta_2^\top)$, $\theta_1 = (\tau_0^\top, \tau_1^\top)^\top$ e $\theta_2 = (\beta^\top, \phi)^\top$. A função de verossimilhança de θ é da forma

$$L(\theta) = \prod_{t=1}^{n} \text{beinf}^*(y_t; \mu_t, \phi, \delta_{0t}, \delta_{1t}) = L_1(\theta_1) L_2(\theta_2),$$

em que

$$\begin{aligned} L_1(\theta_1) &= \prod_{t=1}^{n} \delta_{0t}^{I_{\{0\}}(y_t)} \delta_{1t}^{I_{\{1\}}(y_t)} (1 - \delta_{0t} - \delta_{1t})^{1 - I_{\{0\}}(y_t) - I_{\{1\}}(y_t)}, \\ L_2(\theta_2) &= \prod_{t: y_t \in (0,1)} f(y_t; \mu_t, \phi). \end{aligned}$$

O logaritmo da função de verossimilhança do referido modelo é

$$\begin{aligned} l(\theta) &= \log \left(\prod_{t=1}^{n} \text{beinf}^*(y_t; \mu_t, \phi, \delta_{0t}, \delta_{1t}) \right) \\ &= \sum_{t=1}^{n} \log \left(\text{beinf}^*(y_t; \mu_t, \phi, \delta_{0t}, \delta_{1t}) \right) \\ &= l_1(\theta_1) l_2(\theta_2) \end{aligned}$$

em que

$$l_1(\theta_1) = \sum_{t=1}^{n} l_t(\delta_{0t}, \delta_{1t}),$$
$$l_2(\theta_2) = \sum_{t:y_t\in(0,1)} l_t(\mu_t,\phi),$$

com

$$\begin{aligned} l_t(\mu_t,\phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_t\phi) - \log\Gamma((1-\mu_t)\phi) + (\mu_t\phi - 1)\log(y_t) \\ &\quad + [(1-\mu_t)\phi - 1]\log(1-y_t), \\ l_t(\delta_{0t},\delta_{1t}) &= I_{\{0\}}(y_t)\log(\delta_{0t}) + I_{\{1\}}(y_t)\log(\delta_{1t}) \\ &\quad + (1 - I_{\{0\}}(y_t) - I_{\{1\}}(y_t))\log(1-\delta_{0t}-\delta_{1t}). \end{aligned} \tag{3.3.4}$$

Note que a forma obtida para $l_t(\mu_t,\phi)$ é idêntica ao logaritmo da função de verossimilhança sob o modelo de regressão beta com precisão constante (FERRARI; CRIBARI-NETO, 2004) expresso em (3.1.2). Isso ocorre devido ao fato de que do componente usado para modelar a distribuição beta inflacionada para os valores da variável resposta restritos ao intervalo contínuo (0,1) é o mesmo utilizado naquele modelo. Quando o modelo de regressão BIZU é especificado considerando a precisão variável, a expressão assumirá a forma em (3.2.3), referente ao modelo de regressão beta com precisão variável (SMITHSON; VERKUILEN, 2006).

Os vetores escore são obtidos diferenciando as expressões em (3.3.4) em relação a cada um dos parâmetros desconhecidos. Seja $i = 1, \ldots, k$, os vetores escore para β e ϕ, $U_\beta(\theta_2)$ e $U_\phi(\theta_2)$ possuem entradas

$$\begin{aligned} U_{\beta_i}(\theta_2) &= \frac{\partial l_1(\theta_2)}{\partial \beta_i} \\ &= \sum_{t:y_t\in(0,1)} \phi({y_t}^* - {\mu_t}^*)\frac{1}{g_\mu^{-1}(\mu_t)}x_{ti} \\ &= \sum_{t=1}^{n} \phi[I_{\{0,1\}}(y_t)]({y_t}^* - {\mu_t}^*)\frac{1}{g_\mu^{-1}(\mu_t)}x_{ti}, \\ U_\phi(\theta_2) &= \frac{\partial l_1(\theta_2)}{\partial \phi} \\ &= \sum_{t:y_t\in(0,1)} \mu_t({y_t}^* - {\mu_t}^*) + s(y_t) + \psi(\phi) - \psi((1-\mu_t)\phi) \\ &= \sum_{t=1}^{n}(I_{\{0,1\}}(y_t))\mu_t[{y_t}^* - {\mu_t}^*] + s(y_t) + \psi(\phi) - \psi((1-\mu_t)\phi). \end{aligned}$$

Considerando $L = \text{diag}\{1 - I_{\{c\}}(y_1), \ldots, 1 - I_{\{c\}}(y_n)\}$, então o vetor escore $U_\beta(\theta_2)$

pode ser representada na forma matricial como

$$U_\beta(\theta_2) = \phi X^\top LT(y^* - \mu^*).$$

Para obtenção do vetor escore para τ_0 e τ_1, definimos os vetores $y_{\{0\}} = (I_{\{0\}}(y_1), \ldots, I_{\{0\}}(y_n))^\top$, $y_{\{1\}} = (I_{\{1\}}(y_1), \ldots, I_{\{1\}}(y_n))^\top$, $y_{(0,1)} = (I_{\{0,1\}}(y_1), \ldots, I_{\{0,1\}}(y_n))^\top$, e as matrizes diagonais de dimensão $n \times n$

$$\begin{aligned}
\Delta_0 &= \text{diag}\{1/\delta_{01}, \ldots, 1/\delta_{0n}\},\\
\Delta_1 &= \text{diag}\{1/\delta_{11}, \ldots, 1/\delta_{1n}\},\\
\Delta_{(0,1)} &= \text{diag}\{1/(1-\delta_{01}-\delta_{11}), \ldots, 1/(1-\delta_{0n}-\delta_{1n})\},\\
T_0 &= \text{diag}\{\partial\delta_{01}/\partial\zeta_{01}, \ldots, \partial\delta_{0n}/\partial\zeta_{0n}\},\\
T_1 &= \text{diag}\{\partial\delta_{11}/\partial\zeta_{11}, \ldots, \partial\delta_{1n}/\partial\zeta_{1n}\},\\
T_{01} &= \text{diag}\{\partial\delta_{01}/\partial\zeta_{11}, \ldots, \partial\delta_{0n}/\partial\zeta_{1n}\},\\
T_{10} &= \text{diag}\{\partial\delta_{11}/\partial\zeta_{01}, \ldots, \partial\delta_{1n}/\partial\zeta_{1n}\}.
\end{aligned}$$

Assim, o vetor escore para θ_1 é dado por

$$U(\theta_1) = (U_{\tau_0}(\theta_1)),U_{\tau_1}(\theta_1))^\top,$$

em que

$$\begin{aligned}
U_{\tau_0}(\tau_0,\tau_1) &= {W_0}^\top T_0(\Delta_0 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}) + {W_0}^\top T_{10}(\Delta_1 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}),\\
U_{\tau_1}(\tau_0,\tau_1) &= {W_1}^\top T_{01}(\Delta_0 y_{\{0\}} - \Delta_{(0,1)} y_{(0,1)}) + {W_1}^\top T_1(\Delta_1 y_{\{1\}} - \Delta_{(0,1)} y_{(0,1)}),
\end{aligned}$$

sendo W_0 uma matriz com dimensão $n \times r_0$ em que cada coluna de W_0 representa os n valores observados para a j-ésima covariável de τ_0, $j = 1,2,\ldots,r_0$, e W_1 uma matriz de dimensão $n \times r_1$ em que cada coluna de W_1 representa os n valores observados das covariáveis de τ_1.

O estimador de máxima verossimilhança de $\theta = (\tau_0^\top, \tau_1^\top, \beta^\top, \phi)^\top$ é obtido por meio da resolução do sistema de equações

$$\begin{aligned}
U_{\tau_0}(\theta_1) &= 0,\\
U_{\tau_1}(\theta_1) &= 0,\\
U_\beta(\theta_2) &= 0,\\
U_\phi(\theta_2) &= 0,
\end{aligned}$$

em relação a θ. De forma análoga, não existe resolução analítica fechada para o referido estimador de máxima verossimilhança.

Sob condições usuais de regularidade, a distribuição conjunta aproximada do estimador de máxima verossimilhança de θ é normal $(r_0 + r_1 + k + 1)$-variada, ou seja,

$$\hat{\theta} = (\hat{\tau_0}, \hat{\tau_1}, \hat{\beta}, \hat{\phi})^\top \overset{a}{\sim} N_{r_0+r_1+k+1}\left((\tau_0^\top, \tau_1^\top, \beta^\top, \phi)^\top; K^{***-1}\right),$$

em que K^{***} é a Matriz de Informação de Fisher que assume a forma

$$K^{***} = K^{***}(\theta) = \begin{bmatrix} K_{\tau_0\tau_1}(\theta_1) & 0 \\ 0 & K_{\beta\phi}(\theta_2) \end{bmatrix},$$

onde $K_{\tau_0\tau_1}(\theta_1) = K_{\tau_0\tau_1}$ e $K_{\beta\gamma}(\theta_2) = K_{\beta,\phi}$ são as matrizes de informação de Fisher para $(\tau_0^\top, \tau_1^\top)^\top$ e $(\beta^\top, \phi)^\top$, respectivamente (OSPINA, 2008).

4 Estimação intervalar e testes de hipóteses

4.1 Intervalos de confiança

Na Seção 3 especificamos com detalhes os modelos de regressão beta objetos do presente estudo. Foi visto que a variabilidade assintótica dos estimadores de máxima verossimilhança é dimensionada utilizando a propriedade de normalidade assintótica desses estimadores. Assim, para amostras grandes, e sob condições usuais de regularidade, tal propriedade assegura que a distribuição conjunta aproximada do vetor de estimadores dos parâmetros (θ) é Normal s-variada com

$$\hat{\theta} \overset{a}{\sim} N_s\left(\theta; K^{-1}(\theta)\right),$$

em que θ representa o vetor de parâmetros de interesse, s é a dimensão de θ e K^{-1} é a respectiva a matriz de informação de Fisher contendo as variâncias e covariâncias assintóticas dos estimadores desses parâmetros (OSPINA, 2008). Com isso, para a s-ésima componente de $\hat{\theta}$, $\hat{\theta}_s$, obtemos que

$$\left(\hat{\theta}_s - \theta_s\right)\sqrt{K(\theta)^{ss}} \overset{a}{\sim} N(0,1),$$

em que $K(\theta)^{ss}$ é o (s,s)-ésimo elemento da inversa da matriz de informação de Fisher.

Com esse resultado, pode-se construir intervalos de confiança aproximados para cada parâmetro do respectivo modelo de regressão beta. Desse modo, considerando um nível de confiança de $100(1-\alpha)\%$, um intervalo de confiança aproximado para a s-ésima componente do vetor de parâmetros θ é

$$\left(\hat{\theta}_s - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{K(\theta)}^{ss}};\ \ \hat{\theta}_s + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{K(\theta)}^{ss}}\right), \tag{4.1.1}$$

em que $z_{1-\frac{\alpha}{2}}$ representa o quantil da distribuição Normal padrão tal que $P(Z \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha/2$, com $Z \sim N(0,1)$.

Ospina (2008) ressalta que este resultado permite a construção de intervalos de confiança aproximados para outras grandezas relacionadas aos modelos. Por exemplo, podemos construir um intervalo de confiança para a resposta média μ_t, e para a respectiva razão de chances quando for utilizada a função de ligação logit.

4.2 Testes de hipóteses

Além das estimações pontuais e intervalares discutidas, também é importante avaliar se os coeficientes de regressão são, de fato, estatísticamente significantes. Em outras palavras pretende-se avaliar, considerando a amostra em estudo, se as covariáveis são relevantes para explicar o comportamento da variável resposta. Nesse trabalho, apresentaremos os testes da razão de verossimilhanças e o teste de Wald, que baseiam-se na normalidade assintótica dos estimadores de máxima verossimilhança dos parâmetros de interesse.

Sejam $\beta_1 = (\beta_1, ..., \beta_m)^\top$ e $\beta_1^{(0)} = (\beta_1^{(0)}, ..., \beta_m^{(0)})^\top$, $m \leq k$, e ${\beta_1}^{(0)}$ um vetor de valores dados, consideraremos as hipóteses $H_0 : \beta_1 = \beta_1^{(0)}$ contra $H_1 : \beta_1 \neq \beta_1^{(0)}$. Mais especificamente, temos interesse em $\beta_1^{(0)} = 0$.

4.2.1 Teste da razão de verossimilhanças

Ferrari e Cribari-Neto (2004) descrevem o teste da razão de verossimilhanças (TRV) considerando o modelo de regressão beta com precisão constante discutido na Subseção 3.1, com intuito de verificar se os coeficientes do modelo são significantes, isto é, se estes são diferentes de zero. A estatística da razão de verossimilhanças é dada por

$$\text{TRV} = 2\left\{l(\hat{\theta}) - l(\tilde{\theta})\right\},$$

em que $l(\hat{\theta})$ é o logaritmo da função de verossimilhança avaliado em $\hat{\theta} = (\hat{\beta}^\top, \hat{\phi})^\top$, e $\tilde{\theta} = (\tilde{\beta}^\top, \tilde{\phi})^\top$ é o estimador de máxima verossimilhança de $(\beta^\top, \phi)^\top$ obtido sob a hipótese nula. Observa-se que a estatística TRV calcula o dobro da diferença entre os valores dos logaritmos das funções de verossimilhança sob $\hat{\theta}$ e $\tilde{\theta}$, respectivamente. Desse modo, valores pequenos da estatística indicam não haver diferenças entre o modelo avaliado em $\hat{\theta}$ e o modelo avaliado sob a hipotese nula (BUSE, 1982).

Sob H_0 e condições usuais de regularidade, mostra-se que a estatística do teste possui distribuição aproximada qui-quadrado com m graus de liberdade ($\text{TRV} \overset{a}{\sim} \chi^2_m$). Nesse sentido, considerando um nível de significância α, rejeitamos H_0 em favor de H_1 quando a estatística TRV for maior do que o quantil de ordem $(1-\alpha)$ da distribuição χ^2_m, ou seja, se $\text{TRV} \geq \chi^2_{m,1-\alpha}$, com $\text{P}(\chi^2_m \leq \chi^2_{m,1-\alpha}) = 1-\alpha$.

Da mesma forma, pode-se estender as definições da estatística da razão de verossimilhanças aos demais modelos de regressão beta tratados nas Seções 3.2 e 3.3, como detalhado por Espinheira (2007) e Ospina (2008).

4.2.2 Teste de Wald

Alternativamente ao teste Razão de Verossimilhanças, o teste de Wald (WALD, 1943) também é apresentado por Ferrari e Cribari-Neto (2004) para avaliar a significância dos parâmetros do modelo. Em comparação ao TRV, o teste de Wald é consideravelmente menos conservador uma vez que, para um nível de significância α fixado, observa-se uma taxa de rejeição de H_0 maior do que α (ESPINHEIRA, 2007).

Seja $\hat{\beta}_1$ o estimador de máxima verossimilhança de β_1, então a estatística de Wald é da forma

$$\mathrm{TW} = (\hat{\beta}_1 - \beta_1^{(0)})^\top (\hat{K}_{11}^{\beta\beta})^{-1} (\hat{\beta}_1 - \beta_1^{(0)}),$$

em que $\hat{K}_{11}^{\beta\beta}$ é o (1,1)-ésimo elemento da matriz $K^{\beta\beta}$ definida em (3.1), avaliado em $(\hat{\beta}^\top, \hat{\phi}^\top)^\top$.

Observa-se que a estatística TW baseia-se em uma medida de distância entre o parâmetro sob a hipótese nula $\beta_1^{(0)}$ e a estimativa de máxima verossimilhança de β_1. Assim, tende-se a rejeitar H_0 quanto maior for o valor obtido para a estatística TW.

Assim como a estatística do TRV, a estatística TW também possui distribuição assintótica qui-quadrado com m graus de liberdade ($\mathrm{TW} \overset{a}{\sim} \chi^2_m$), e rejeitamos H_0 sempre que, dado um nível de significância α, estatística TW for maior do que o quantil de ordem $(1-\alpha)$ da distribuição χ^2_m, ou seja, se $\mathrm{TW} \geq \chi^2_{m,1-\alpha}$.

Para os casos em que tivermos apenas um parâmetro de regressão, então β será um escalar e a estatística de Wald se reduzirá à expressão

$$\mathrm{TW} = \frac{(\hat{\beta} - \beta^{(0)})^2}{\widehat{\mathrm{Var}(\hat{\beta})}},$$

com $\mathrm{TW} \overset{a}{\sim} \chi^2_1$. Observe que, nesse caso, a estatística TW equivale a z^2 em que

$$z = \sqrt{\mathrm{TW}} = \frac{\hat{\beta} - \beta^{(0)}}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta})}}},$$

com $z \overset{a}{\sim} N(0,1)$, o que nos permite a realização de teste de hipóteses equivalente ao teste de Wald, porém utilizando a distribuição normal padrão.

Espinheira (2007) e Ospina (2008) apresentam a estatística de Wald para os modelos de regressão beta com precisão variável e para os modelos inflacionados, respectivamente.

5 Critérios de seleção de modelos

No processo de modelagem é comum nos depararmos com situações onde temos mais de um modelo que, em uma primeira análise, se mostre adequado para descrever a variabilidade da variável resposta a partir das variáveis explicativas disponíveis. Nessas situações torna-se importante utilizar ferramentas para a seleção do modelo mais adequado.

Essa tarefa é mais simples quando os modelos em comparação são modelos encaixados. Dizemos que dois modelos são encaixados quando um deles é obtido a partir do outro, de maior complexidade, impondo algum tipo de restrição aos parâmentros (CORDEIRO; DEMÉTRIO, 2008). Usualmente, essa restrição pode ser feita igualando um ou mais parâmetros do modelo mais complexo a zero ou a outra constante cujo valor se pretenda testar. Nessas situações, podem ser utilizados um dos testes de hipóteses apresentados na Seção 4.

Adicionalmente aos testes de hipóteses discutidos na Subseção 4.2, foram desenvolvidas outras técnicas com o propósito de obter critérios objetivos que possam ser usados na seleção de modelos, a exemplo dos critérios de informação (HASTIE; TIBSHIRANI, 1990). Nesse estudo serão utilizados o critério de informação de Akaike (AIC) (AKAIKE, 1974), o critério de informação bayesiano (BIC) (SCHWARZ, 1978) e o critério consistente de informação de Akaike (CAIC) (AKAIKE, 1983), que são definidos, respectivamente, por

$$\begin{aligned} \text{AIC} &= -2l(\hat{\theta}) + 2d, \\ \text{BIC} &= -2l(\hat{\theta}) + 2\log(n), \\ \text{CAIC} &= -2l(\hat{\theta}) + 2[\log(n) + 1], \end{aligned}$$

em que em que $l(\hat{\theta})$ é o logaritmo da função de verossimilhança avaliado em $\hat{\theta} = (\hat{\beta}^\top, \hat{\phi})^\top$, n o número de realizações da variável resposta, e d o número de parâmetros do modelo.

Conforme observado por Ospina (2008), esses critérios baseiam-se em uma penalização da função de verossimilhança na medida em que o modelo se torna mais complexo, ou seja, aumenta o número de parâmetros. Assim, modelos com maiores valores para a função de verossimilhança, ou seja, com melhor qualidade de ajuste, tendem a ser bonificados, enquanto, por outro lado, são penalizados a cada parâmetro existente no modelo. Nesse sentido, quanto maior for a medida para os critérios elencados, maior será a informação perdida e, sendo assim, será selecionado o modelo que apresente o menor valor para o critério de informação utilizado.

6 Técnicas de diagnóstico

Após o ajuste do modelo é importante efetuar uma avaliação diagnóstica com o intuito de verificar se as suposições assumidas previamente permanecem válidas, além de verificar a qualidade do ajuste obtido para os dados. Nesse sentido, a análise de resíduos pode ajudar nessa avaliação.

Em um modelo de regressão, os resíduos ordinários podem ser definidos como sendo a diferença entre o valor observado da variável resposta y e o correspondente valor ajustado pelo modelo (CHARNET et al., 1999). Assim, podemos dizer que os resíduos medem a discrepância entre o modelo ajustado e os valores observados para a variável resposta y no conjunto de dados amostrais.

6.1 Resíduos ponderados padronizados

Para os modelos de regressão beta apresentados nas Subseções 3.1 e 3.2, serão considerados os resíduos padronizados ponderados 2. Para isso, definimos inicialmente o resíduo ponderado r_t^*, que segundo Espinheira (2007), considerando o modelo de regressão beta com precisão constante, é da forma

$$r_t^* = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\phi v_t}}, \tag{6.1.1}$$

em que $v_t = \mathrm{Var}(y_t^*) = \psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)$.

Espinheira (2007) propõe uma padronização da medida em (6.1.1) obtendo o resíduo ponderado padronizado 2 (r_t^{pp}) como sendo

$$r_t^{pp} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{v_t(1-h_{tt}^*)}}, \tag{6.1.2}$$

em que h_{tt}^* o t-ésimo elemento da diagonal principal da matriz H^*, expressa como

$$H^* = \widehat{W}^{\frac{1}{2}} X (X^\top \widehat{W} X)^{-1} X^\top \widehat{W}^{\frac{1}{2}},$$

em que as matrizes W e X foram definidas na Subseção 3.1.

Espinheira (2007) apresenta o resíduo ponderado padronizado 2 para o modelo de regressão beta com precisão variável, adaptando a expressão em (6.1.2) por meio de

$$r_t^{pp} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\hat{v}_t(1-\hat{h}_{tt}^*)}} \tag{6.1.3}$$

em que, para esse caso, $\hat{v}_t = \psi'(\hat{\mu}_t\hat{\phi}_t)+\psi'((1-\hat{\mu}_t)\hat{\phi}_t)$ e $\hat{h}^*_{tt}$ é o t-ésimo elemento da diagonal principal da matriz

$$\widehat{H}^* = (\widehat{W}\Phi)^{\frac{1}{2}}X(X^\top\hat{\Phi}\widehat{W}X)^{-1}X^\top(\widehat{W}\Phi)^{\frac{1}{2}},$$

em que Φ é a matriz definida na Subseção 3.2.

6.2 Resíduos quantis aleatorizados

Para os modelos de regressão beta inflacionados apresentados na Subseção 3.3 serão considerados os resíduos quantis aleatorizados. Conforme mencionado por Ospina (2008), o resíduo quantil aleatorizado foi definido por Dunn e Smyth (1996) e pode ser considerado uma versão aleatorizada do resíduo de Cox & Snell (COX; SNELL, 1968).

Definimos o resíduo quantil aleatorizado para os modelos de regressão beta inflacionados como

$$r_t^q = \Phi^{-1}(u_t), \tag{6.2.1}$$

em que $t = 1, \ldots, n$, $\Phi(.)$ denota a fda da distribuição normal padrão, e u_t é uma variável aleatória uniforme no intervalo $(a_t, b_t]$, sendo que a_t e b_t são definidos conforme modelo de regressão utilizado.

Para o modelo de regressão beta inflacionado em zero ou em um (BEZI ou BEOI) $a_t = \lim_{y\uparrow y_t}\text{BI}_c(y, \hat{\alpha}, \hat{\mu}, \hat{\phi})$ e $b_t = \text{BI}_c(y, \hat{\alpha}, \hat{\mu}, \hat{\phi})$, em que $\text{BI}_c(y, \alpha, \mu, \phi)$ é a fda da distribuição beta inflacionada em zero ou em um definida em (2.2.3), e $\hat{\alpha}$, $\hat{\mu}$ e $\hat{\phi}$ são os estimadores de máxima verossimilhança de α, μ e ϕ, respectivamente.

Conforme observa Ospina (2008), no caso do modelo de regressão BEZI, temos que para $y_t = 0$, $a_t = 0$, $b_t = \hat{\alpha}$ e, logo, u_t possui distribuição uniforme no intervalo $(0, \hat{\alpha}]$. Para o modelos BEOI, com $y_t = 1$, temos que a_t e b_t se reduzem para $(1 - \hat{\alpha})$ e 1, respectivamente, e, portanto, u_t passa a ser distribuída segundo uma distribuição uniforme no intervalo $(1 - \hat{\alpha}, 1]$.

Para o caso do modelos de regressão BEINF com a precisão ϕ constante, os resíduos quantis aleatorizados seguem a mesma expressão definida em (6.2.1), porém, sendo u_t uma variável aleatória uniformemente distribuída no intervalo $(a_t, b_t]$, temos que $a_t = \lim_{y\uparrow y_t}\text{BEINF}(y, \pi_t, \gamma_t, \mu_t, \phi)$ e $b_t = \text{BEINF}(y_t, \pi_t, \gamma_t, \mu_t, \phi)$, em que $\text{BEINF}(y, \pi, \gamma, \mu, \phi)$ é a fda da distribuição BEINF definida em (2.2.7).

6.3 Envelope simulado

Um método gráfico relevante na avaliação da veracidade do pressuposto referente à distribuição de probabilidade assumida para a variável resposta é o envelope simulado,

introduzido por Atkinson (1985).

Segundo Fernandes (2019), o envelope simulado é uma técnica que consiste na inclusão, em um gráfico normal de probabilidades, de bandas obtidas por meio de amostras geradas pelo método de Monte Carlo a partir do modelo ajustado. Tais bandas fornecem um referencial para a flutuação dos pontos, auxiliando na identificação de possíveis afastamentos entre valores realizados da variável resposta e a distribuição de probabilidades teórica assumida.

Considerando uma amostra de n realizações para a variável resposta, tomamos $t_{[i]}$, $i = 1,2,\ldots,n$, os valores ordenados de um resíduo de interesse, a exemplo do resíduo ponderado padronizado 2. Conforme Fernandes (2019), o gráfico normal de probabilidades contém os pontos $(\mathrm{E}(Z_{[i]}), t_{[i]})$, em que $\mathrm{E}(Z_{[i]})$ representa os valores esperados das estatísticas de ordem da distribuição normal padrão que, segundo Blom (1958), podem ser aproximados por

$$\mathrm{E}(Z_{[i]}) \approx \Phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right).$$

Para determinação dos limites das bandas e construção do envelope empírico, Atkinson (1985) sugere os passos a seguir:

Passo 1. Ajustar o modelo a partir da amostra original e calcular os n resíduos de interesse (resíduos originais).

Passo 2. Gerar n observações simuladas (réplicas) a partir do modelo ajustado no Passo 1.

Passo 3. Ajustar novamente o modelo utilizando as n observações geradas no Passo 2.

Passo 4. Calcular os n resíduos (réplicas) de interesse com base no modelo gerado no Passo 3.

Passo 5. Repetir os Passos 2 a 4 m vezes.

Passo 6. Arranjar cada um dos m grupos de n resíduos em ordem crescente de valor.

Passo 7. Calcular os valores máximos e mínimos para cada um dos m grupos de resíduos obtidos, que serão, respectivanente, os limites superiores e inferiores do envelope. Assim, para cada resíduo haverá um limite inferior e superior calculado empiricamente. Também, podem-se definir percentis de interesse para os limites das bandas do envelope, como, por exemplo, os referentes a 5% e 95%.

Segundo Atkinson (1985), o objetivo da simulação é fornecer amostras de resíduos com mesma estrutura de covariância do modelo ajustado. Além disso, o autor sugere usar $m = 19$, de modo que a probabilidade do maior dos resíduos exceder o limite superior do envelope seja próximo a 1/20.

A Figura 6 contém um exemplo de gráfico de probabilidade contendo o envelope simulado, representado pela área entre os limites das duas linhas que formam a banda em torno dos pontos. Observa-se que, caso uma proporção considerável de pontos esteja fora dos limites do envelope, então teremos indícios de que o modelo ajustado pode não ser adequado.

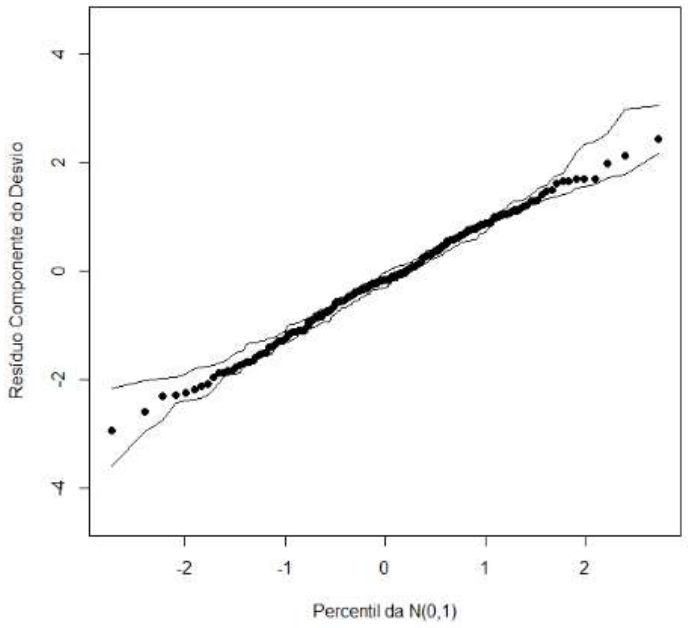

Figura 6 Exemplo de envelope simulado. (FERNANDES, 2019)

Entretanto ressalta-se que, embora o método do envelope simulado forneça um indício importante e de fácil interpretação, este deve ser usado com cautela uma vez que inexiste critério objetivo para definição da quantidade ou proporção de pontos fora do envelope para fins de rejeição ou não de um modelo. Assim, este método deve ser utilizado em conjunto com outros disponíveis, a exemplo dos discutidos anteriormente.

6.4 *Worm plot*

Proposto por Buuren e Fredriks (2001), o *worm plot* é outra ferramenta gráfica de diagnóstico bastante útil para análise de resíduos de modelos de regressão. Este método permite identificar o quanto um modelo de regressão se ajusta ao conjunto de dados em

estudo e, além disso, em quais locais o ajuste pode ser melhorado (BUUREN, 2007).

Stasinopoulos, Rigby e Bastiani (2018) explicam que o *worm plot* consiste em uma sequência de gráficos do tipo *QQ-Plot* sem tendência com curvas elípticas que indicam bandas com confiança aproximada de 95%. Segundo Buuren e Fredriks (2001) o eixo vertical do gráfico contém, para cada realização da variável resposta, a diferença entre as medidas de locação teóricas e empíricas (*deviation*) e no eixo horizontal são alocados os quantis teóricos da distribuição normal padrão, conforme ilustrado na Figura 7.

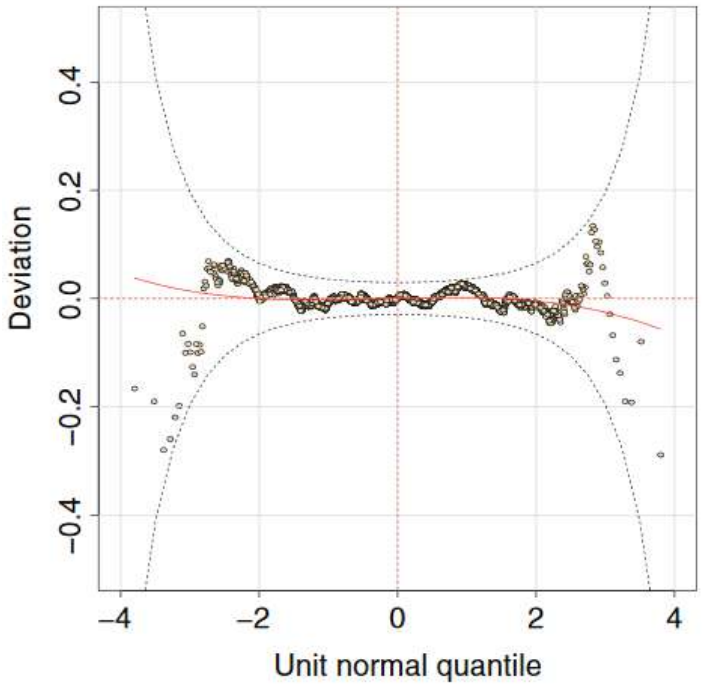

Figura 7 Exemplo de *worm plot*. (STASINOPOULOS; RIGBY; BASTIANI, 2018)

Assim, para um modelo bem ajustado espera-se que os pontos no gráfico sigam próximos à linha horizontal no meio do gráfico, sem comportamento sistemático e, em sua maioria, dentro dos limites das bandas de confiança. Stasinopoulos, Rigby e Bastiani (2018) consideram que o modelo avaliado no *worm plot* da Figura 7 é aceitável, uma vez que inexistem pontos fora dos limites das bandas de confiança.

Buuren e Fredriks (2001) esclarecem interpretações sobre a média, variância, assimetria e curtose considerando diversos comportamentos da curva obtida em um gráfico *worm plot*, conforme elencado na Tabela 1.

Tabela 1: Interpretações de alguns momentos com base no comportamentos da curva obtida com o *worm plot*. (BUUREN; FREDRIKS, 2001).

Forma	Momento	Comportamento da curva	Conclusão
Intercepto	Média	Acima da origem	Média é subestimada
		Abaixo da origem	Média é superestimada
Inclinação	Variância	Inclinação positiva	Variância é subestimada
		Inclinação negativa	Variância é superestimada
Parábola	Assimetria	Forma de U	Excesso de assimetria à esquerda
		Forma de U invertido	Excesso de assimetria à direita
Curva em S	Curtose	Forma de S crescente	Cauda é excessivamente leve
		Forma de S decrescente	Cauda é excessivamente pesada

Assim como outros métodos gráficos, o diagnóstico por meio do *worm plot* deve ser utilizado e interpretado em conjunto com outras técnicas, uma vez que não é possível definir de forma objetiva um critério para rejeição ou aceitação de um modelo de regressão específico.

7 Metodologia

7.1 Métodos

O presente Relatório Final foi desenvolvido por meio de estudos dos modelos de regressão beta introduzidos por Ferrari e Cribari-Neto (2004), Smithson e Verkuilen (2006) e Ospina (2008). Inicialmente, foram investigadas as principais características associadas a cada uma dessas técnicas para identificar as vantagens e limitações de sua utilização. Posteriormente, os modelos estudados foram aplicados a dados reais e também a dados simulados, onde tivemos oportunidade de verificar de forma prática as características identificadas no estudo téorico.

7.2 Apoio computacional

Todos os cálculos e avaliações numéricas relacionadas às estimações dos parâmetros dos modelos, bem como os gráficos gerados ao longo desse trabalho, foram realizados com suporte computacional utilizando a linguagem de programação R e o software estatístico de mesmo nome, em sua versão 4.1.0. O software R (R Core Team, 2022) é de domínio público e está disponível gratuitamente para download no endereço eletrônico http://www.r-project.org/.

7.2.1 Biblioteca betareg

Para ajuste dos modelos e análises relacionadas à regressão beta com precisão constante (FERRARI; CRIBARI-NETO, 2004) e precisão variável (SMITHSON; VERKUILEN, 2006) foi utilizada a biblioteca `betareg` (CRIBARI-NETO; ZEILEIS, 2010) do software R. Esse recurso permite obter ajustes para o modelo de regressão beta quando a variável resposta está restrita ao intervalo contínuo aberto (0,1). A referida implementação utiliza a forma reparametrizada da distribuição beta, definida em (2.1.4), e a formulação dos modelos de regressão apresentados em (3.1.1), para o caso da precisão constante, e em (3.2.1), no caso do modelo com precisão variável.

A estimação dos parâmetros que representam a média, aqui denotada por μ e a precisão, aqui denotada por ϕ, é efetuada por meio do método da máxima verossimilhança apresentado nas Subseções 3.1 ou 3.2, conforme o caso.

7.2.2 Biblioteca GAMLSS

Para os casos em que a variável resposta assume valores em [0,1), (0,1] ou [0,1], será utilizada a biblioteca `GAMLSS` (STASINOPOULOS; RIGBY; STASINOPOULOS, 2006), também do software R. Esta biblioteca contém a implementação de modelos de regressão e técnicas de diagnóstico baseadas em diversas distribuições de probabilidade, dentre as quais estão a distribuição beta inflacionada em zero ou um, e a distribuição beta inflacionada em zero e em um (OSPINA; FERRARI, 2010).

A biblioteca `GAMLSS` utiliza as distribuições beta inflacionadas com parametrizações diferentes das apresentadas nas Subseções 2.2.1 e 2.2.2. Para a distribuição beta inflacionada em zero ou um, a partir da fdp em (2.2.1), é utilizada a parametrização

$$\begin{aligned} \mu &= \mu, \\ \sigma &= \sqrt{\frac{1}{\phi+1}}, \\ \nu &= \frac{\alpha}{1-\alpha}, \end{aligned} \tag{7.2.1}$$

resultando na fdp da forma

$$bi_c{}^*(y;\nu,\mu,\sigma) = \begin{cases} \frac{\nu}{1+\nu}, & \text{se } \; y = c, \\ \left(\frac{1}{1+\nu}\right) f(y,\mu,\sigma), & \text{se } \; 0 < y < 1, \end{cases} \tag{7.2.2}$$

em que c representa a probabilidade de observar zero ($c = 0$) ou um ($c = 1$), $\nu \in (0,\infty)$ e $f(y;\mu,\sigma)$ é a fdp da distribuição beta na forma

$$f(y;\mu,\sigma) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\frac{\mu(1-\sigma^2)}{\sigma^2}\right)\Gamma\left(\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)} y^{\frac{\mu(1-\sigma^2)}{\sigma^2}-1}(1-y)^{\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}-1}, \;\; 0 < y < 1, \tag{7.2.3}$$

com $\mu \in (0,1)$ e $\sigma \in (0,1)$.

Observa-se que o novo parâmetro σ tem uma interpretação inversa à de ϕ, uma vez que representa a dispersão dos dados, enquanto o parâmetro ν representa a chance (*odds*) de ocorrência do valor extremo c. Ainda, como $0 < \sigma < 1$ ($\phi > 0$) e $\nu > 0$ ($0 < \alpha < 1$), temos que as funções de ligação adequadas para σ são as mesmas utilizadas para μ, e as ligações para ν são aquelas adequadas para ϕ.

Para a distribuição BEINF, a partir da fdp em (2.2.8), tomamos

$$\begin{aligned} \mu &= \mu, \\ \sigma &= \sqrt{\frac{1}{\phi+1}}, \\ \omega_0 &= \frac{\delta_0}{1-\delta_0-\delta_1}, \\ \omega_1 &= \frac{\delta_1}{1-\delta_0-\delta_1}, \end{aligned} \tag{7.2.4}$$

que resulta na fdp

$$\text{beinf}^{**}(y;\delta_0,\delta_1,\mu,\phi) = \begin{cases} \frac{\omega_0}{1+\omega_0+\omega_1}, & \text{se } \ y=0, \\ \frac{\omega_1}{1+\omega_0+\omega_1}, & \text{se } \ y=1, \\ \frac{1}{1+\omega_0+\omega_1} f(y;\mu,\sigma), & \text{se } \ 0<y<1, \end{cases} \tag{7.2.5}$$

sendo $f(y;\mu,\sigma)$ a fdp da distribuição beta na forma apresentada em (7.2.3), $\mu,\sigma \in (0,1)$ e $\omega_0, \omega_1 \in (0,\infty)$. Observa-se que, para esse caso, as ligações adequadas para ω_0 e ω_1 são as mesmas utilizadas para ϕ.

Para os dois casos, a técnica de estimação dos parâmetros dos modelos inflacionados utilizado na biblioteca `GAMLSS` é o método da máxima verossimilhança.

8 Resultados e discussões

Efetuada a revisão bibliográfica, os modelos de regressão beta estudados foram aplicados a dados reais e também a dados simulados. Tais ajustes tiveram por objetivo enfatizar as características e propriedades identificadas no estudo teórico, e servirão para ressaltar as vantagens e limitações de cada modelo, considerando as diferentes situações observadas nos dados utilizados.

8.1 Estudos de simulação

Adicionalmente às aplicações que serão discutidas na Subseção 8.2, foram realizados estudos de simulação com o objetivo de melhor avaliar as propriedades dos modelos de regressão beta, bem como enfatizar as vantagens de utilização de cada um nos diferentes cenários apresentados.

Todas as simulações foram realizadas utilizando o software R e foram baseadas em 5.000 réplicas de Monte Carlo para tamanhos amostrais n de 40, 80, 160 e 320. Para cada réplica é gerada uma amostra aleatória da variável resposta, $y = (y_1, \ldots, y_n)^\top$ considerando a distribuição referente ao cenário apresentado. As matrizes de regressão contendo as covariáveis foram extraídas, também de forma aleatória, da distribuição uniforme padrão ($\mathcal{U}(0,1)$), e foram mantidas constantes ao longo de todas as réplicas do respectivo tamanho amostral.

Para fins de avaliação dos resultados e comparação entre os ajustes, foram computadas as estimativas dos vieses relativos (VMR) e dos erros quadráticos médios relativos (EQMR) para cada tamanho amostral, e para cada parâmetro fixado. Além disso, estimamos os erro quadrático médio (EQM) dos valores ajustados para a resposta. As fórmulas são dadas por

$$\begin{aligned}
\text{VMR} &= \frac{1}{R}\sum_{i=1}^{R}\left(\frac{\widehat{\theta}_i - \theta}{\theta}\right), \\
\text{EQMR} &= \frac{1}{R}\sum_{i=1}^{R}\left(\frac{\widehat{\theta}_i - \theta}{\theta}\right)^2, \\
\text{EQM} &= \frac{1}{R}\sum_{i=1}^{R}\left[\frac{1}{n}\sum_{t=1}^{n}\left(\widehat{y}_t - y_t\right)^2\right],
\end{aligned}$$

em que R é a quantidade de réplicas de Monte Carlo geradas, e $\widehat{\theta}_i$ é a estimativa do parâmetro θ na i-ésima réplica. Além disso, foram utilizadas representações gráficas dessas grandezas para auxiliar na visualização dos resultados.

8.1.1 Cenário 1: Regressão beta sem inflação

Para esse cenário foram consideradas estruturas de regressão para a média μ e para a precisão ϕ dadas por

$$\begin{aligned} g_\mu(\mu_t) &= \beta_1 + \beta_2 x_{t1} + \beta_3 x_{t2} = X_t^\top \beta, \\ g_\phi(\phi_t) &= \gamma_1 + \gamma_2 z_{t1} + \gamma_3 z_{t2} = Z_t^\top \gamma, \end{aligned} \tag{8.1.1}$$

em que X_t é vetor contendo os valores das covariáveis associadas à média μ_t e Z_t é o vetor de covariáveis para a precisão ϕ_t. Para $g_\mu(\cdot)$ e $g_\phi(\cdot)$ foram utilizadas as funções de ligação logit e logarítmica, respectivamente, e foram fixados os valores $\beta_1 = 1{,}4$, $\beta_2 = 1{,}0$, $\beta_3 = 1{,}0$, $\gamma_1 = 1{,}0$, $\gamma_2 = 0{,}7$ e $\gamma_3 = 0{,}7$, de modo que, para as amostras geradas, as medianas de μ e ϕ ficaram próximas a 0,92 e 5,40, respectivamente.

O experimento consistiu nos seguintes passos:

Passo 1. Gerar as amostras considerando o modelo de regressão beta acima especificado.

Passo 2. Para cada réplica, ajustou-se um modelo de regressão beta com precisão variável, tido como o modelo correto, utilizando as mesmas covariáveis usadas para geração das amostras.

Passo 3. Para as mesmas réplicas, ajustou-se um modelo de regressão beta com precisão constante, ou seja, fazendo $\gamma_2 = \gamma_3 = 0$ em (8.1.1), tido como o modelo incorreto. Para a estrutura de regressão da média, foram utilizadas as mesmas covariáveis usadas na estrutura análoga do modelo gerador dos dados.

Passo 4. Efetuou-se uma comparação dos resultados dos modelos descritos nos Passos 2 e 3.

A Tabela 2 apresenta os respectivos VMRs e EQMRs dos coeficientes associados aos submodelos da média e precisão. Conforme esperado, os VMRs sob o modelo correto são relativamente pequenos e tendem a zero conforme se aumenta o tamanho amostral. A mesma análise sob o modelo incorretamente especificado aponta que os VRMs dos coeficientes do submodelo da média, além de estarem em patamares absolutos superiores aos verificados sob o modelo correto, não parecem reduzir com o aumento do tamanho amostral, o que sugere a existência vieses. Quanto a γ_1 no modelo incorreto, observa-se que, também conforme esperado, apresenta um valor relativamente distante do valor verdadeiro. Analisando os EQMRs, vemos que a percepção quanto à adequabilidade das estimativas dos coeficientes no modelo correto é reforçada, uma vez que estes valores são relativamente pequenos, e tendem a reduzir conforme se eleva o tamanho amostral. Sob

o modelo incorreto, os EQMRs das estimativas dos coeficientes são relativamente bem superiores aos do modelo correto.

Tabela 2: Cenário 1 - VMRs e EQMRs obtidos para as estimativas dos coeficientes de regressão segundo tamanho amostral n e modelo aplicado. Os EQMRs estão entre parênteses.

n	Modelo	β_1	β_2	β_3	γ_1	γ_2	γ_3
40	Correto	0,03138	0,02126	0,03589	0,16517	0,03389	0,05331
		(0,19081)	(0,79192)	(0,74063)	(0,60206)	(2,47012)	(2,35724)
	Incorreto	0,27182	-0,40583	-0,40259	0,73614	-	-
		(0,35851)	(0,52766)	(0,47645)	(0,62310)	(-)	(-)
80	Correto	0,02002	0,01692	0,00303	0,09905	0,01981	-0,02597
		(0,07409)	(0,29532)	(0,35997)	(0,21890)	(0,93750)	(1,08596)
	Incorreto	0,24758	-0,4117	-0,39517	0,65750	-	-
		(0,22232)	(0,30240)	(0,33004)	(0,47148)	(-)	(-)
160	Correto	0,01462	0,00596	-0,01319	0,05683	-0,00455	-0,02816
		(0,04718)	(0,15863)	(0,17470)	(0,13177)	(0,47122)	(0,49900)
	Incorreto	0,27149	-0,41141	-0,42777	0,67132	-	-
		(0,20581)	(0,24798)	(0,26419)	(0,47015)	(-)	(-)
320	Correto	0,00383	0,00240	0,00235	0,01720	0,01371	-0,00116
		(0,02085)	(0,07873)	(0,07978)	(0,05677)	(0,21928)	(0,22805)
	Incorreto	0,26170	-0,43021	-0,41878	0,63569	-	-
		(0,16306)	(0,22211)	(0,21690)	(0,41383)	(-)	(-)

Na Figura 8 estão dispostos os boxplots para os valores das estimativas dos parâmetros associados à média μ para os dois modelos. Por meio da imagem é possível observar o comportamento discutido na análise dos VMRs e EQMRs, uma vez que as estimativas sob o modelo correto estão centradas em torno dos valores verdadeiros e a precisão aumenta para amostras maiores. Além disso, identifica-se de forma visual os vieses introduzidos nas estimativas dos parâmetros sob a especificação incorreta. Mais especificamente, as estimativas sob o modelo incorretamente especificado estão centradas em torno de valores errados para os respectivos parâmetros. Considerando que, geralmente, a interpretabilidade é um dos objetivos do processo de modelagem, os vieses introduzidos nas estimativas dos parâmetros associados ao submodelo da média, poderiam trazer significativos prejuízos à correta interpretação do impacto de cada covariável na média da variável resposta.

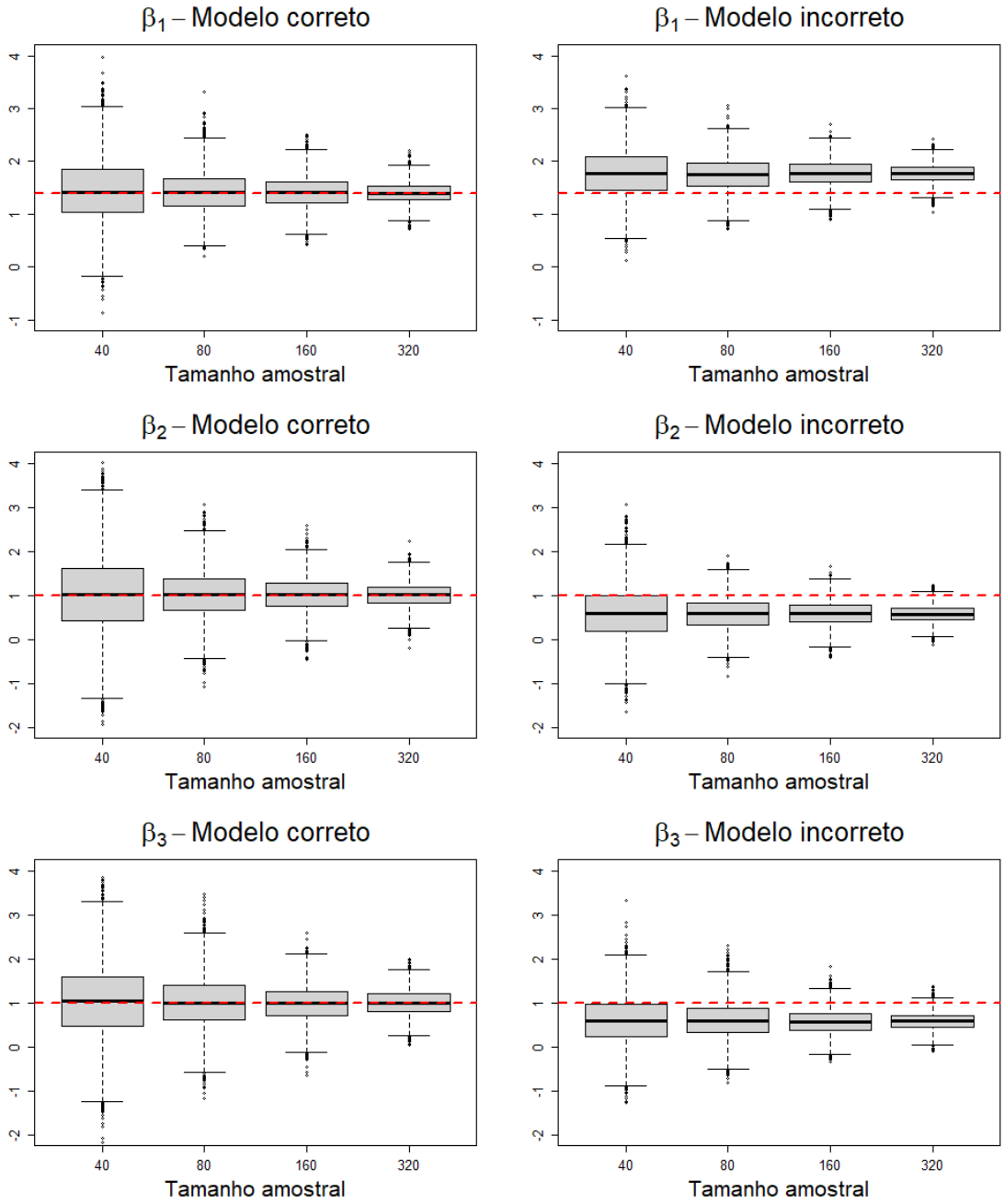

Figura 8 Boxplots das estimativas dos parâmetros de μ sob o Cenário 1 segundo o modelo aplicado. As linhas vermelhas tracejadas representam os valores reais dos parâmetros.

Na Figura 9 estão dispostos os boxplots das estimativas dos coeficientes associados ao submodelo da precisão. Assim como ocorre no submodelo da média, as estimativas dos coeficientes sob a especificação correta estão centradas em torno dos valores verdadeiros e tendem a ser mais precisas conforme se eleva o tamanho amostral. Além disso, vemos que sob o modelo incorreto a estimativa de γ_1 fica distante do valor real.

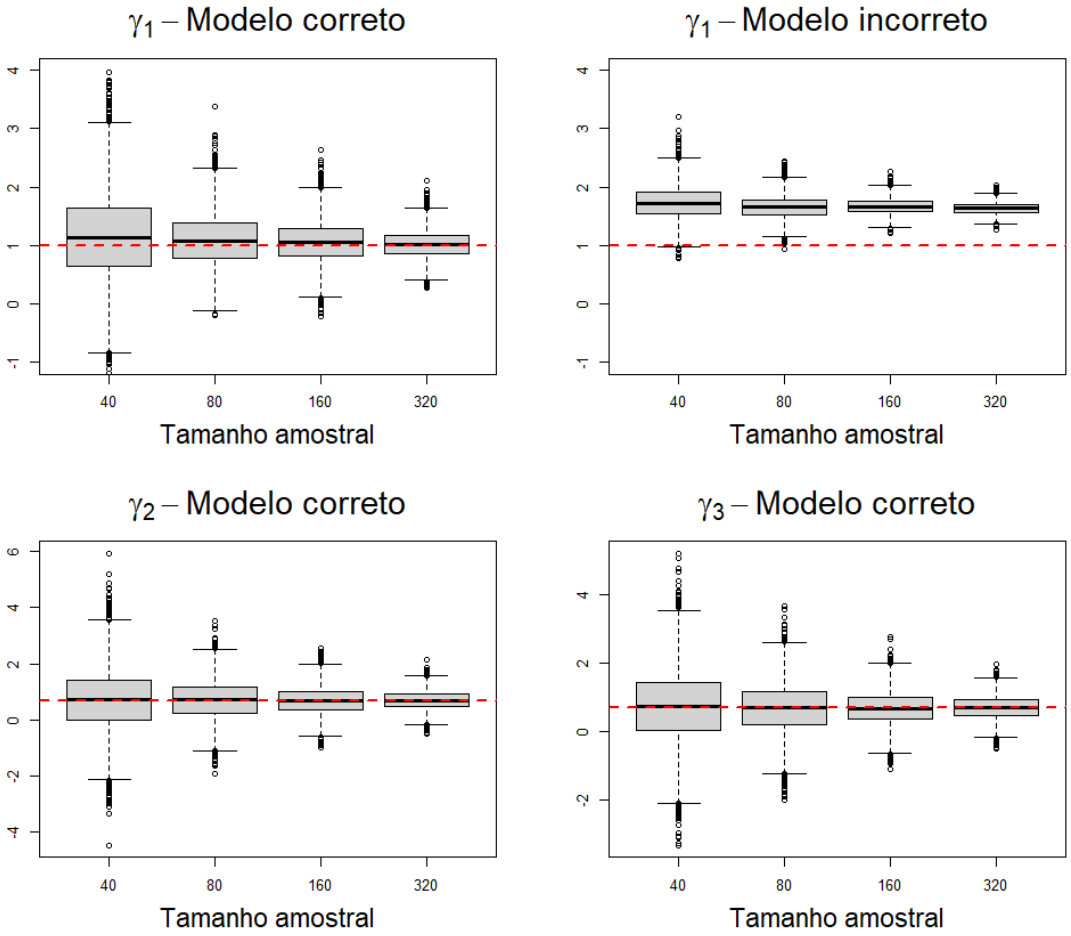

Figura 9 Boxplots das estimativas dos parâmetros de ϕ sob o Cenário 1 segundo o modelo aplicado. As linhas vermelhas tracejadas representam os valores reais dos parâmetros.

A Tabela 3 mostra os EQMs para os valores preditos pelos modelos, que são os valores ajustados de μ. Observa-se que, mesmo sob o modelo incorreto, os EQMs se mantém em patamares baixos para todos os tamanhos amostrais, além de serem bem próximos aos EQMs obtidos sob o modelo correto. Isso sugere que, apesar dos vieses introduzidos nas estimativas dos coeficientes, o modelo incorretamente especificado parece ter se ajustado bem aos dados.

Tabela 3: Cenário 1 - EQMs segundo tamanho amostral n e modelo aplicado.

n	40	80	160	320
Modelo correto	0,01305	0,01381	0,01290	0,01351
Modelo incorreto	0,01336	0,01415	0,01314	0,01375

Portanto, dadas as especificações simuladas no presente cenário, mostra-se a van-

tangem de se utilizar a regressão beta com precisão variável quando esta característica estiver presente nos dados de interesse. Outro ponto interessante também identificado foi referente aos valores preditos pelo modelo incorreto, que não ficaram tão distantes dos valores verdadeiros, apesar dos vieses nas estimativas dos coeficientes de regressão. Isso pode ser explicado pelo fato do modelo de regressão beta, mesmo com a precisão tratada como constante, ser heterocedástico. Essa propriedade decorre da variância de y_t ser função não somente de ϕ_t, mas também de μ_t. Assim, ainda que a precisão não seja modelada, a estrutura de regressão usada na modelagem da média acaba sendo naturalmente acomodada na expressão da variância, dando flexibilidade ao ajuste.

Considerando as especificações simuladas no presente cenário, o modelo com precisão constante, mais parcimonioso, poderia ser uma alternativa ao modelo com precisão variável, caso o objetivo seja, por exemplo, somente obter os valores ajustados de μ para fins de predição, uma vez que não ocorreram diferenças significantes nos valores ajustados.

8.1.2 Cenário 2: Regressão beta inflacionada em zero e em um

Conforme mencionado na Seção 7, para os modelos de regressão BEINF discutidos na Subseção 3.3.2 está sendo considerada as distribuições sob a parametrização disponível na biblioteca `GAMLSS` (STASINOPOULOS; RIGBY; STASINOPOULOS, 2006), que é a expressa em (7.2.4). Assim, as estruturas de regressão consideradas para esse cenário são

$$\begin{aligned} g_\mu(\mu_t) &= \beta_1 + \beta_2 x_{t1} = X_t^\top \beta, \\ g_\sigma(\sigma) &= \gamma, \\ g_{\omega_0}(\omega_{0t}) &= \tau_{01} + \tau_{02} w_{0t1} = W_{0t}^\top \tau_0, \\ g_{\omega_1}(\omega_{1t}) &= \tau_{11} + \tau_{12} w_{1t1} = W_{1t}^\top \tau_1, \end{aligned} \tag{8.1.2}$$

em que W_{0t} e W_{1t} são os vetores de regressão associados à ω_{0t} e ω_{1t}, respectivamente. Observe que o modelo está sendo especificado considerando a dispersão σ como constante ao longo das observações. A função logit foi utilizada como ligação para $g_\mu(\cdot)$ e $g_\sigma(\cdot)$, a função logarítmica foi utilizada como ligação em $g_{\omega_0}(\cdot)$ e $g_{\omega_1}(\cdot)$ e os valores dos parâmetros foram $\beta = (-1{,}0, 2{,}0)^\top$, $\gamma = 0{,}2$, $\tau_0 = (-1{,}4, 1{,}4)^\top$ e $\tau_1 = (0{,}7, -1{,}4)^\top$. Com isso, para as amostras geradas, obtém-se as medianas de μ, σ, ω_0 e ω_1 próximas a 0,50, 0,55, 0,49 e 1,01, respectivamente.

O experimento consistiu nos passos:

Passo 1. Gerar as amostras considerando o modelo de regressão BEINF acima especificado.

Passo 2. Para cada réplica de Monte Carlo, ajustou-se um modelo de regressão

BEINF com precisão constante, tido como o correto, utilizando, nas três estruturas de regressão, as mesmas covariáveis usadas para geração das amostras.

Passo 3. As mesmas réplicas do passo anterior foram ajustadas para alterar os valores 0 e 1 da variável resposta para 0,001 e 0,999, respectivamente.

Passo 4. Para cada réplica alterada conforme passo anterior, ajustou-se um modelo de regressão beta com precisão constante e sem inflação, tido como o incorreto. Para as estruturas de regressão associadas à média e à precisão foram utilizadas as mesmas covariáveis usadas nas estruturas análogas do modelo gerador dos dados.

Passo 5. Efetuou-se uma comparação dos resultados dos modelos descritos nos Passos 2 e 4.

Na Tabela 4 estão dispostos os respectivos VMRs e EQMRs dos coeficientes associados aos parâmetros dos modelos ajustados.

Tabela 4: Cenário 2 - VMRs e EQMRs obtidos para as estimativas dos coeficientes de regressão segundo tamanho amostral n e modelo aplicado. Os EQMRs estão entre parênteses.

n	Modelo	β_1	β_2	γ_1	τ_{01}	τ_{02}	τ_{11}	τ_{12}
40	Correto	0,03012	0,02989	-0,79225	0,21352	0,24292	0,07866	0,08379
		(0,48263)	(0,31777)	(2,72797)	(1,26938)	(2,31468)	(1,43511)	(1,12560)
	Incorreto	-1,72974	-1,43513	6,55404	-	-	-	-
		(3,13626)	(2,16192)	(43,24637)	(-)	(-)	(-)	(-)
80	Correto	0,01160	0,01336	-0,37311	0,07666	0,08571	0,0476,	0,04215
		(0,18090)	(0,12080)	(0,96068)	(0,37035)	(0,75060)	(0,53533)	(0,42402)
	Incorreto	-1,71988	-1,42878	6,64752	-	-	-	-
		(3,02103)	(2,08606)	(44,31601)	(-)	(-)	(-)	(-)
160	Correto	0,00775	0,00549	-0,18254	0,03343	0,03931	0,01546	0,01733
		(0,08272)	(0,06871)	(0,43679)	(0,14428)	(0,37579)	(0,24342)	(0,24298)
	Incorreto	-1,70562	-1,42154	6,70122	-	-	-	-
		(2,93869)	(2,04205)	(44,96674)	(-)	(-)	(-)	(-)
320	Correto	0,00428	0,00306	-0,07996	0,01421	0,01601	0,01311	0,01236
		(0,04666)	(0,03421)	(0,19108)	(0,07400)	(0,17512)	(0,13784)	(0,11778)
	Incorreto	-1,71286	-1,42498	6,69930	-	-	-	-
		(2,94880)	(2,04304)	(44,91097)	(-)	(-)	(-)	(-)

Como era esperado, os VMRs sob o modelo correto são relativamente pequenos e tendem a se aproximar de zero quando se eleva o tamanho da amostra. Observa-se que, sob o modelo correto e no maior tamanho amostral, todos os coeficientes apresentaram vieses médios relativos próximos de 1%, exceto γ_1, cujo valor obtdo foi de -8%. Ao

analisarmos o modelo incorreto, percebemos que os VRMs apresentaram valores relativos muito superiores aos verificados para o modelo correto, e não tendem a reduzir quando se aumenta o tamanho amostral.

De forma análoga, a avaliação dos EQMRs confirma a percepção inicial quanto à adequabilidade das estimativas sob o modelo corretamente especificado, desde que os vieses observados são relativamente pequenos e tendem a diminuir quando aumenta-se o tamanho da amostra. Sob o modelo incorretamente especificado, observou-se EQMRs em patamares relativamente muito superiores aos obtidos sob o modelo correto. Além disso, vemos que estes erros não mostram uma tendência de redução com o aumento do tamanho das amostras.

Na Figura 10 estão dispostos os boxplots para os valores das estimativas dos coeficientes de regressão associados a μ e ϕ no modelo incorretamente especificado. Na imagem é possível visualizar os vieses introduzidos nas estimativas quando se utilizou a especificação incorreta. Conforme havia sido identificado por meio dos VRMs e EQMs, os parâmetros β_1 e γ_1 foram superestimados, enquanto que o parâmetro β_2 foi subestimado.

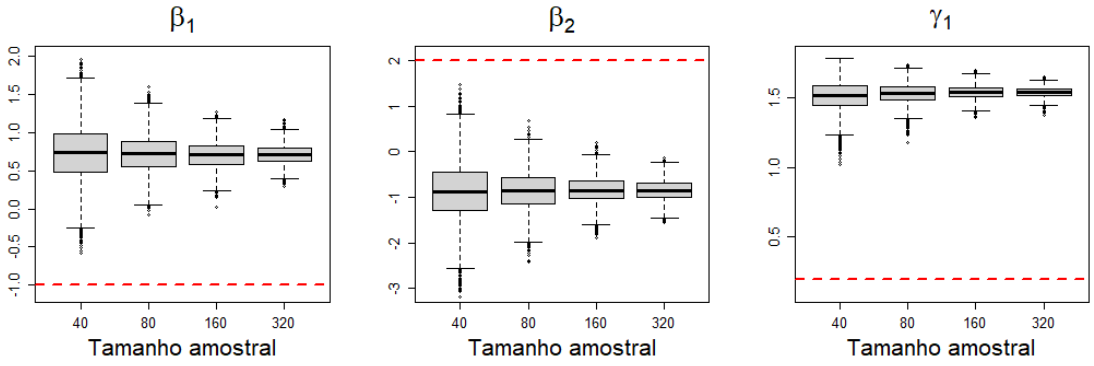

Figura 10 Boxplots das estimativas dos parâmetros sob o Cenário 2 utilizando o modelo com a especificação incorreta. As linhas vermelhas tracejadas representam os valores reais dos parâmetros.

A Figura 11 ilustra os boxplots das estimativas dos coeficientes de regressão sob o modelo correto. Verifica-se que, à exceção de γ_1, as medianas das estimativas dos coeficientes ficaram muito próximas aos valores verdadeiros dos parâmetros, inclusive nos menores tamanhos amostrais. Além disso é possível identificar uma tendência de aumento na precisão destas estimativas à medida que são utilizadas amostras maiores. Em relação a γ_1, conforme foi identificado por meio da análise dos VMRs e EQMRs, o viés obtido foi, em termos relativos, consideravelmente maior do que os verificados para os demais coeficientes, especialmente para $n = 40$.

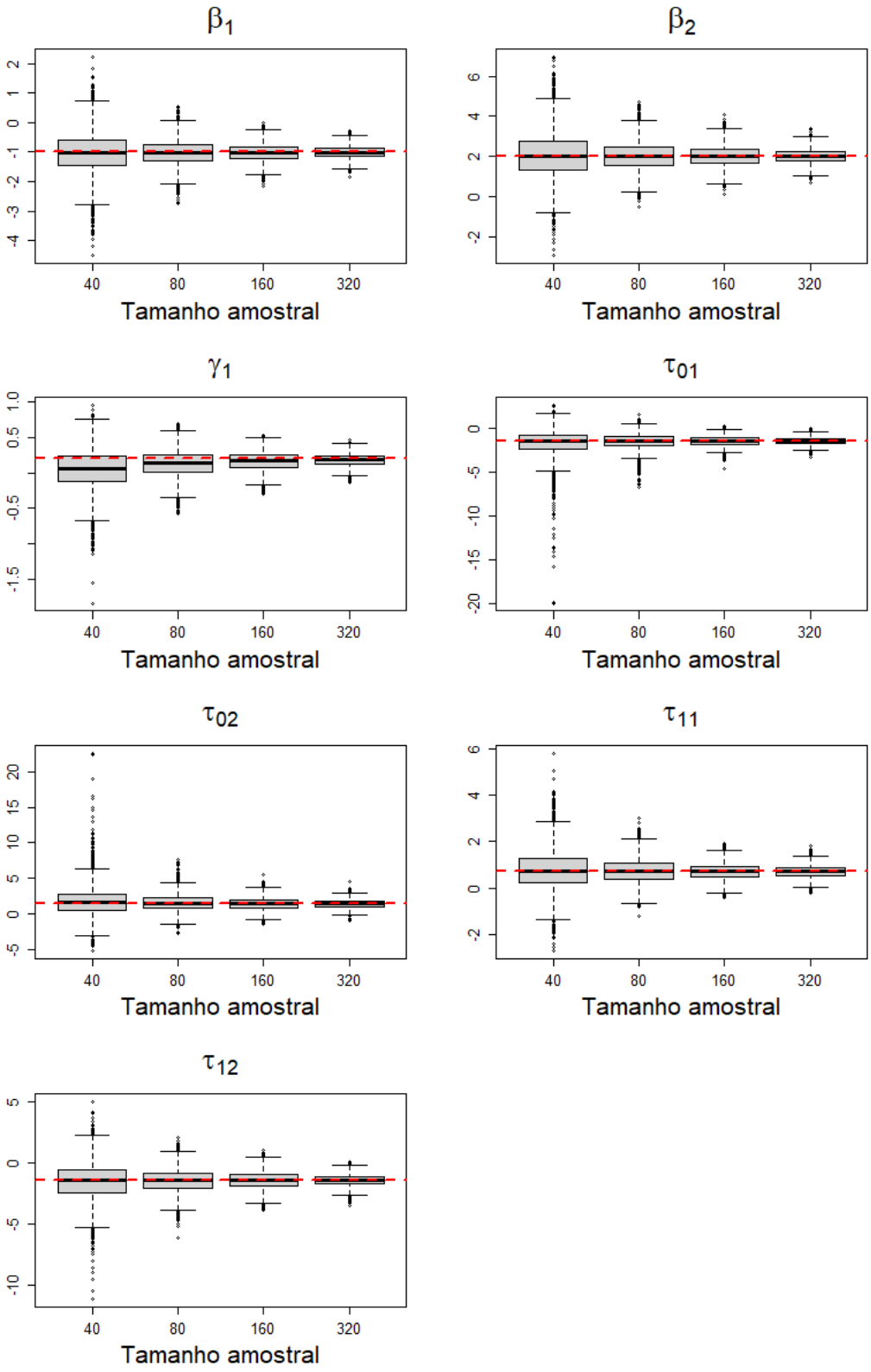

Figura 11 Boxplots das estimativas dos parâmetros sob o Cenário 2 utilizando o modelo com a especificação correta. As linhas vermelhas tracejadas representam os valores reais dos parâmetros.

Observa-se que, sob o modelo incorreto, a média da variável resposta é determinada pela estrutura de regressão associada ao parâmetro μ. Assim, estes vieses verificados nas estimativas dos coeficientes sob o modelo incorreto podem trazer prejuízos à uma adequada estimação dos valores preditos, bem como enviesar a interpretação da relação entre as covariáveis e a variável resposta.

Portanto, observou-se que os melhores resultados são obtidos ajustando os dados originais utilizando o modelo BEINF, que é o método adequado para esse tipo de situação. Desse modo, considerando as condições do experimento, pode-se dizer que não é recomendada a alteração dos valores 0 e 1 na variável resposta para trazê-los ao intervalo contínuo (0,1) e, desse modo, viabilizar a utilização da regressão beta convencional.

8.2 Aplicações

Conforme será detalhado adiante, os dados utilizados nas aplicações foram extraídos de trabalhos acadêmicos e artigos sobre o assunto, e foram analisados de acordo com a base teórica apresentada nos capítulos anteriores.

Salvo quando estiver expressamente informado, os resíduos utilizados nas análises de diagnóstico são os resíduos ponderados padronizados 2, quando envolver a regressão beta sem inflação, e os resíduos quantís aleatorizados, para os modelos de regressão beta inflacionados. Para construção das bandas de confiança dos envelopes simulados nos gráficos de probabilidade normal, estão sendo considerados os percentis referentes a 5% e 95% dos resíduos simulados. Além disso, em todos os testes de hipóteses efetuados e intervalos de confiança construídos fixamos nível de significância de 5%.

8.2.1 Aplicação 1: Fator de simultaneidade para sistemas prediais de gás natural

Os dados, analisados anteriormente por Ferrari (2014) e Fernandes (2019), foram coletados a partir de um estudo sobre distribuição de gás natural para utilização em prédios residenciais em São Paulo, Brasil. Conforme Fernandes (2019), o estudo foi conduzido pelo Instituto de Pesquisas Tecnológicas (IPT) e pela Companhia de Gás de São Paulo (COMGÁS), com o objetivo principal de identificar oportunidades de melhoria no dimensionamento da sua rede de distribuição de gás natural.

O conjunto de dados contém 42 observações referentes a medições efetuadas em sistemas prediais de clientes da COMGÁS no ano de 2004, onde foram registradas informações sobre o fator de simultaneidade e a potência computada, e são a base de dados utilizada nessa aplicação. Segundo Fernandes (2019), o fator de simultaneidade, cujos

valores estão contidos no intervalo unitário, é uma informação útil para o correto dimensionamento da rede de distribuição de gás, e a potência computada se refere ao consumo máximo de energia dos eletrodomésticos, em megawatts. Maiores detalhes sobre os dados podem ser obtidos em Fernandes (2019).

O objetivo da aplicação é explicar a relação entre o fator de simultaneidade, que é a variável resposta, e a potência computada. Na Tabela 5 são apresentadas algumas medidas de posição para o fator de simultaneidade. Conforme se observa, o fator de simultaneidade assume valores entre 0,016 e 0,464. Apesar da mediana ser próxima da média, estas medidas estão muito próximas do 1º e 3º quartis e de zero, o que sugere uma concentração de valores próximos a zero e, portanto, distribuição da resposta assimétrica à esquerda.

Tabela 5: Medidas descritivas de posição para o fator de simultaneidade

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,016	0,061	0,070	0,097	0,114	0,464

Como a variável resposta está restrita ao intervalo (0,1), os modelos de regressão beta estudados nas Subseções 3.1 e 3.2 podem ser adequados à natureza do presente problema. Admitindo-se que as realizações $y_1, \ldots, y_n$ da resposta (fator de simultaneidade) são provenientes de variáveis aleatórias independentes tal que cada y_t, $t = 1, \ldots, n$, tem distribuição beta com média μ_t e precisão ϕ, consideramos as estruturas de regressão para a média μ_t e para a precisão ϕ dadas por

$$\begin{aligned} g_\mu(\mu_t) &= \beta_1 + \beta_2 x_t^*, \\ g_\phi(\phi) &= \gamma, \end{aligned} \tag{8.2.1}$$

em que β_1, β_2 e γ são os parâmetros de regressão, e $x_{t_1}^* = \log(x_{t_1})$, sendo x_t^* o vetor de valores conhecidos da variável explicativa (logaritmo da potência computada). Inicialmente, foram ajustados 4 modelos considerando para $g_\mu(\cdot)$ as funções de ligação logit, probit, cloglog e loglog. Em todos os ajustes foi considerada a função logarítmica como ligação em $g_\phi(\cdot)$.

Na Figura 12 estão apresentados os gráficos de probabilidade normal dos resíduos com os envelopes simulados para os modelos ajustados. Para todos os ajustes, verifica-se que os resíduos estão, em sua maioria, concentrados dentro dos limites das bandas do envelope, o que indica que não há afastamentos quanto ao pressuposto de distribuição da variável resposta.

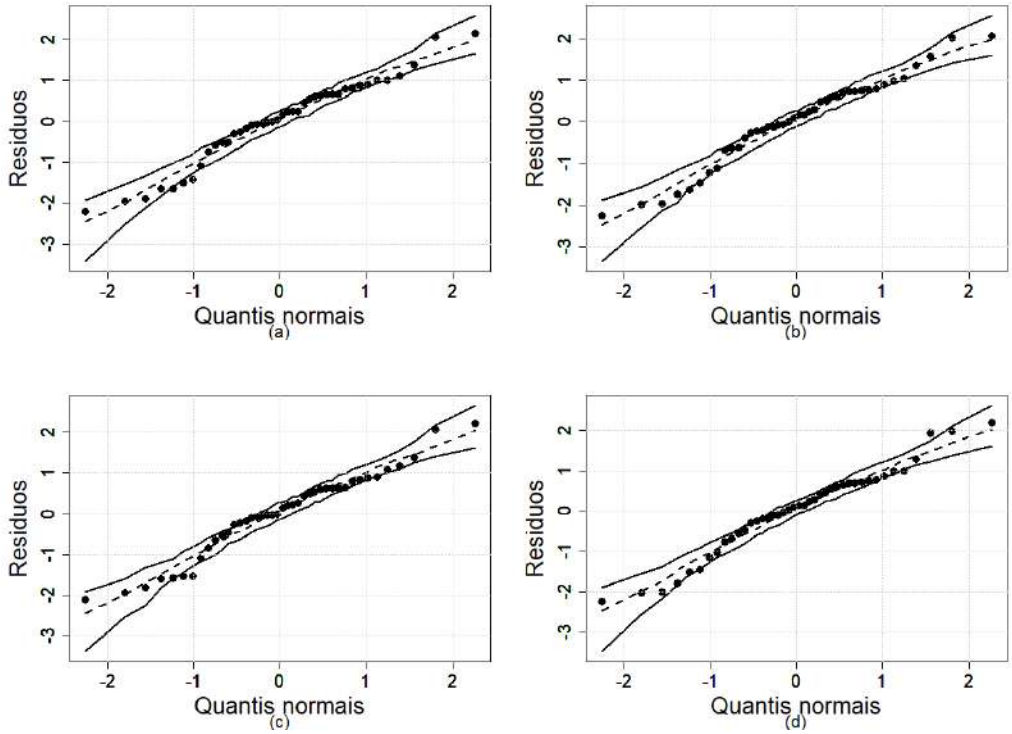

Figura 12 Gráficos de probabilidade normal com envelope simulado para os resíduos dos modelos com precisão constante ajustados com ligação logit (a), probit (b), complementar loglog (c) e loglog (d).

Na Figura 13 consta o gráfico de dispersão entre o fator de simultaneidade e a potência computada com as curvas ajustadas sob os ajustes considerados. Percebe-se que as curvas ajustadas estão muito próximas entre si, como também, parecem modelar bem a relação entre as variáveis.

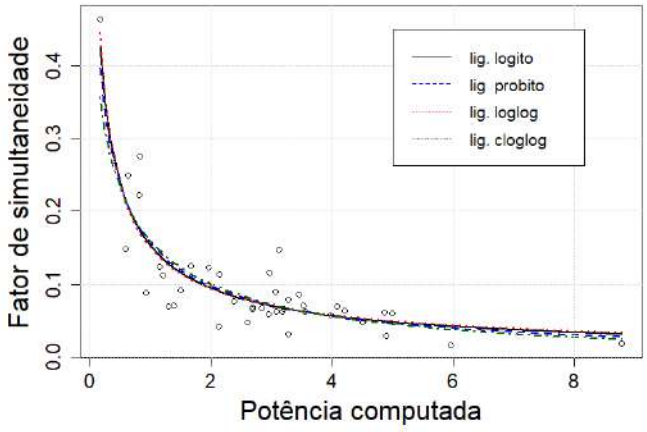

Figura 13 Gráfico de dispersão entre a potência computada e o fator de simultaneidade e as curvas ajustadas pelos modelos considerados.

Adicionalmente, foram ajustados modelos de regressão beta considerando a precisão como variável ao longo da amostra. Para isso, o logaritmo da potência computada foi utilizado como covariável regressora também na precisão ϕ_t. Assim, as estruturas são dadas por

$$\begin{aligned} g_\mu(\mu_t) &= \beta_1 + \beta_2 x^*_{t1}, \\ g_\phi(\phi_t) &= \gamma_1 + \gamma_2 x^*_{t1}, \end{aligned} \tag{8.2.2}$$

em que γ_1, γ_2 são os coeficientes de regressão associados à ϕ_t. Para fins de comparação, foram utilizadas, para $g_\mu(\cdot)$ e $g_\phi(\cdot)$, as mesmas funções de ligação dos modelos definidos em (8.2.1). A Figura 14 apresenta os gráficos normais de probabilidades com os envelopes simulados para cada modelo com precisão variável ajustado. Observa-se que, para todos os modelos ajustados, não houve ganho na qualidade do ajuste em comparação aos modelos com precisão constante ajustados anteriormente.

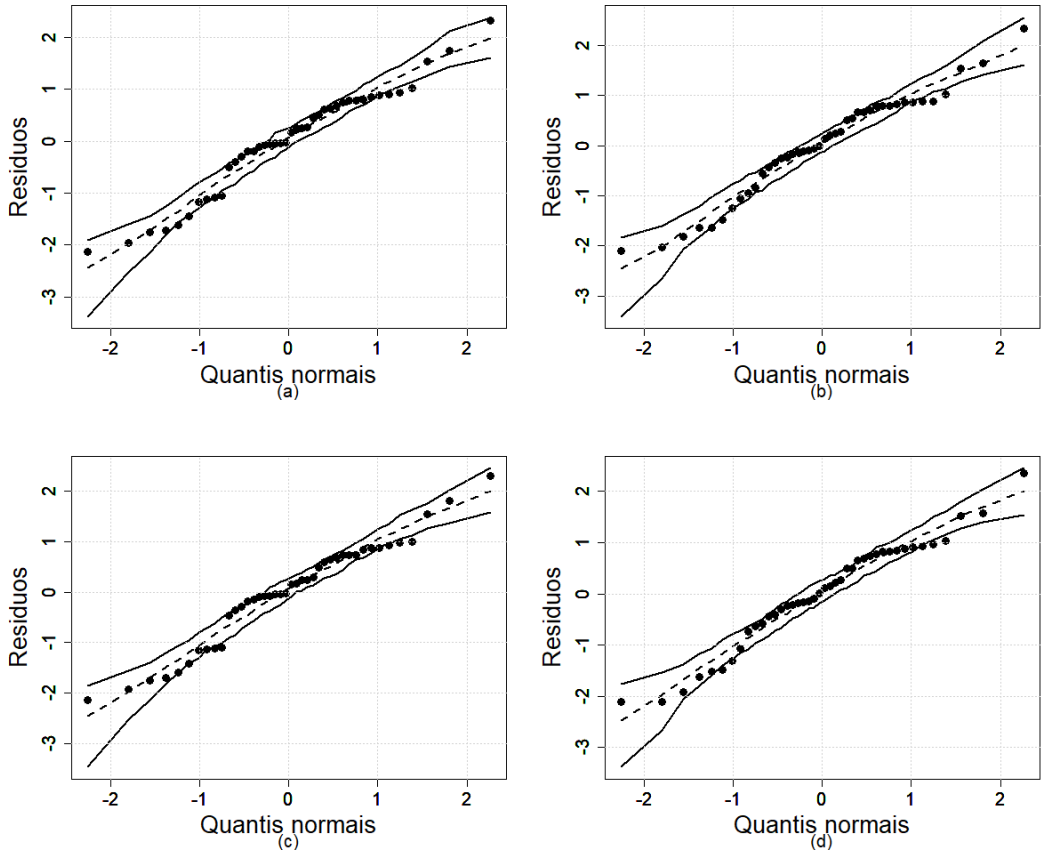

Figura 14 Gráficos de probabilidade normal com envelope simulado para os resíduos dos modelos com precisão variável ajustados com ligação logit (a), probit (b), complementar loglog (c) e loglog (d).

Nesse sentido, seguiremos a análise utilizando o modelo de regressão beta com

precisão constante e ligação logit, sendo este menos complexo, e nos permite uma interpretação em termos da razão de chances, o que pode ser útil para um melhor entendimento da relação entre as variáveis. Portanto, para a ligação escolhida temos que μ_t e ϕ são dados por

$$\mu_t = \frac{e^{\beta_1+\beta_2 x^*_{t1}}}{1+e^{\beta_1+\beta_2 x^*_{t1}}}, \quad \phi = \exp(\gamma). \tag{8.2.3}$$

A Tabela 6 apresenta as estimativas para os parâmetros do modelo selecionado, além dos erros padrão, estatísticas z e p-valores do teste de Wald para a nulidade dos coeficientes da regressão. Observa-se que o logaritmo da potência computada é estatisticamente significante para explicar o comportamento do fator de simultaneidade.

Tabela 6: Estimativas dos coeficientes de regressão, erros padrão, estatísticas z e p-valores do teste de Wald para a nulidade dos coeficientes da regressão beta com precisão constante.

Parâmetro	Estimativa	Erro padrão	Est. z	p-valor
β_1	-1,712	0,067	-25,48	$< 0{,}001$
β_2	-0,794	0,067	-11,93	$< 0{,}001$
γ	4,374	0,219	19,94	$< 0{,}001$

Adicionalmente, realizamos o teste da razão de verossimilhanças para testar se o coeficiente de regressão associado à covariável é diferente de zero, ou seja, testou-se $H_0 : \beta_2 = 0$ contra $H_1 : \beta_2 \neq 0$. O teste apontou evidências de adequabilidade do modelo, uma vez que a estatística TRV apresentou o valor de 49,33, ou seja, TRV $> \chi^2_{1,0,95} = 3{,}84$. Assim, ao nível de significância de 5%, rejeitamos H_0 em favor de H_1.

Para interpretação do modelo, de (8.2.3) temos que

$$\widehat{\mu}_t = \frac{e^{-1{,}71-0{,}79x^*_{t_1}}}{1+e^{-1{,}71-0{,}79x^*_{t_1}}},$$

em que x^* representa o logaritmo da potência computada. Observa-se que a relação é inversa pois $\widehat{\beta}_2$ é negativo, ou seja, conforme se reduz o logaritmo da potência computada, a média do fator de simultaneidade aumenta.

A Tabela 7 elenca, para diferentes valores do logaritmo da potência computada (x^*_{t1}), o impacto percentual na média estimada da variável resposta ao acrescermos 0,1 na covariável, o que representa aproximadamente aumento de 1,1 megawats na potência. Conforme se observa para $x^*_{t1} = 0{,}5$, ao acrescermos 0,1 nesta covariável, a média estimada do fator de simultaneidade aumenta em aproximadamente 6,83%. Esse impacto tende a ser maior quanto maior for o valor do logaritmo da potência computada.

Tabela 7: Impacto na média estimada da variável resposta ao acrescermos 0,1 em $x^*_{t_1}$ segundo diferentes valores da covariável para o modelo de regressão beta com precisão constante e ligação logit.

Valor de $x^*_{t_1}$	Alteração na resposta média (%)
-1,5	4,91
-1,0	5,55
0,5	6,83
1,0	7,06
1,5	7,23
2.0	7,34

Por fim, mesmo chegando a um bom ajuste aos dados utilizando a regressão beta, é importante salientar que, a depender da situação ou contexto do problema a ser resolvido, outras técnicas também podem conduzir a resultados satisfatórios. Para fins comparativos, ajustamos um modelo de regressão linear normal considerando a transformação logit para variável resposta (RLN logit), uma vez que esta foi a ligação utilizada no modelo de regressão beta selecionado. A estrutura de regressão fica dada por

$$y^*_t = g_\mu(y_t) = \log\left(\frac{y_t}{1-y_t}\right) = \beta_1 + \beta_2 x^*_{t1} + \varepsilon_i, \text{ com } \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2), \tag{8.2.4}$$

em que y_t é o fator de simultaneidade para a t-ésima observação e $g_\mu(\cdot)$ é a função logit. Ressalta-se que a variável resposta y^* segue uma distribuição normal com média μ_t e variância σ^2 constante.

A Figura 15 apresenta o gráfico de dispersão entre o fator de simultaneidade e a potência computada juntamente com as curvas ajustadas para o modelo RLN logit, para o modelo final de regressão beta, e para um modelo de RLN ajustado de forma análoga ao RLN logit, porém sem transformação na resposta (RLN). A estrutura de regressão do modelo RLN é data por $y_t = \beta_1 + \beta_2 x^*_{t1} + \varepsilon_i$. Conforme se observa, o modelo RLN não é adequado ao presente problema, uma vez que não aparenta acomodar bem a relação entre os dados, além de ajustar valores negativos, situação incompatível com a característica da resposta. Em contrapartida, o modelo RLN logit aparenta estar se ajustando bem a relação contida nos dados. Isso é reforçado pela proximidade deste com a curva ajustada do modelo de regressão beta selecionado.

Entretanto, mesmo com um ajuste razoável aos dados, temos que a transformação efetuada na resposta para a RLN logit impossibilita a interpretação direta do coeficiente estimado de x^*_t em termos da média estimada do fator de simultaneidade, que também é um passo importante do processo de modelagem. Desse modo, caso o intuito fosse, por exemplo, somente obter a predição do fator de simultaneidade a partir da potência computada, sem maior interesse pela interpretabilidade dessa relação, então o modelo

RLN logit poderia ser uma alternativa a ser considerada, uma vez que é relativamente mais simples do que a regressão beta.

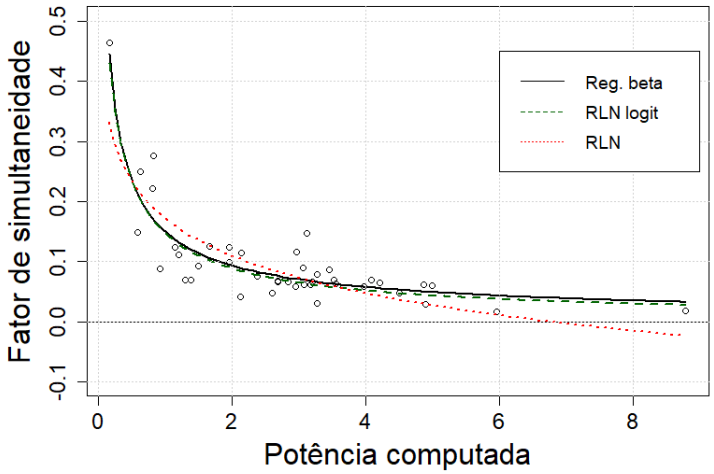

Figura 15 Gráfico de dispersão entre a potência computada e o fator de simultaneidade e as curvas ajustadas para os modelos RLN logit, RLN e regressão beta com precisão constante e ligação logit.

8.2.2 Aplicação 2: Impacto nas capturas de atum devido ao aumento da temperatura do oceano

Nesta aplicação foram utilizados os dados disponibilizados por Monllor-Hurtado, Pennino e Sanchez-Lizaso (2017), cujo estudo objetivou analisar o impacto do aquecimento dos oceanos na pesca global. Segundo os autores, o aquecimento dos oceanos está afetando a pesca no mundo, uma vez que observa-se um aumento nas capturas de espécies de peixes de águas mais quentes em latitudes mais altas, além da redução nas capturas de espécies tropicais e subtropicais em áreas delimitadas pelos trópicos. Isso pode indicar um movimento das populações de peixes em direção aos polos em resposta à elevação das temperaturas dos oceanos.

Segundo Monllor-Hurtado, Pennino e Sanchez-Lizaso (2017), o estudo se concentrou no atum tropical devido ao fato de sua distribuição pelos oceanos ser altamente condicionada à temperatura do mar, o que torna a distribuição da espécie um bom indicador do efeito da mudança climática na pesca global.

Os dados contém observações referentes a 19.019 tentativas individuais de capturas de peixes com um palangre entre 1967 e 2011 nos Oceanos Índico, Pacífico e Atlântico. O palangre, conforme ilustrado na Figura 16, é constituído por uma linha principal, forte e comprida, de onde partem outras linhas secundárias mais curtas, em grande número e

em intervalos regulares, com um anzol ao final de cada uma delas (WIKIPEDIA, 2023).

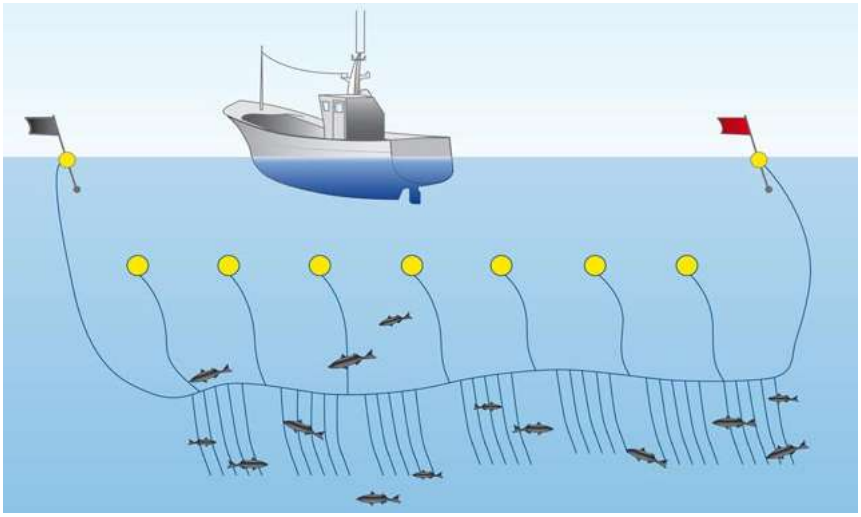

Figura 16 Exemplo de palangre. (ESPESCA, 2023)

Além das informações citadas, os dados contém:

- latitude e longitude aproximada do local da pesca;
- a temperatura da superfície da água do mar (temperatura), em graus Celsius (° C);
- o esforço empreendido (esforço), expresso em número de anzóis/ganchos utilizados;
- proporção de atum tropical entre todos os peixes capturados, que será a variável resposta dessa aplicação.

O objetivo principal da presente aplicação é explicar a relação entre a proporção de atum tropical capturado por meio das demais variáveis disponíveis, em especial em função da temperatura da superfície da água do mar no momento da captura. Aqui, consideramos um subconjunto dos dados. Foram utilizadas somente os registros de pescas ocorridas no oceano Atlântico durante o ano de 1972. Na Tabela 8 apresentamos as principais medidas descritivas para estes dados. Conforme se observa, a variável resposta, que é uma proporção, assume valores no intervalo [0,1), ou seja, incluindo o valor 0. Isso significa que ocorreram tentativas onde não tivemos nenhum atum tropical dentre os peixes capturados. Desse modo, o modelo de regressão BEZI estudado na Subseção 3.3.1 pode ser adequado à natureza do problema.

Tabela 8: Medidas descritivas de posição para a proporção de atum tropical capturado.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,0000	0,0005	0,0237	0,1921	0,1988	0,9733

O subconjunto de dados selecionado possui 140 observações nas quais 35, ou para 25% dos dados, a variável resposta assume o valor zero. Observa-se que o 1º quartil e a mediana, são muito próximas de zero, além da mediana ser consideravelmente menor do que a média. Isso sugere uma distribuição assimétrica à esquerda da resposta, com uma maior concentração de observações que assumem valores próximos a zero. Foram consideradas inicialmente as covariáveis latitude, longitude e temperatura. A variável esforço não será utilizada, uma vez que foram identificadas observações em que esta assume valores que passam dos três milhões, o que nos pareceu incompatível com a sua descrição.

As variáveis latitude e longitude determinam, juntas, o local da pesca. A latitude informa o quão próxima dos polos da terra é a localização, assumindo valores positivos no hemisfério norte, negativos no hemisfério sul e zero sobre a linha do Equador. A longitude diz respeito à distância em relação ao meridiano de Greenwich, onde assume o valor zero, ao longo da linha do equador. No hemisfério oriental, ou seja, à leste do meridiano de Greewich, assume valores positivos e no hemisfério ocidental assume valores negativos. Desse modo, a coordenada $0/0^\circ$ representa o ponto onde o meridiano de Greenwich e a linha do Equador se cruzam (WIKIPEDIA, 2023). Assim, obtivemos que as longitudes observadas no subconjunto de dados variam entre -100° e 20°, e as latitudes entre -45° e 45°, o que é compatível com a localização do oceano Atlântico em relação às coordenadas geográficas.

A análise das coordenadas geográficas dos locais de pesca mostrou indícios de correlação negativa entre a latitude e a variável resposta. Ao tomarmos os valores absolutos dessas latitudes, obtemos, a um nível de 95% de confiança, o intervalo $[-0{,}76; -0{,}57]$ para o coeficiente de correlação de Pearson. O teste de hipóteses para correlação linear (MORETTIN; BUSSAB, 2017) apontou forte significância dessa correlação, com p-valor inferior a 0,001. Esse resultado indica que quanto maior for o valor absoluto da latitude, menor é a proporção de atum tropical capturado. Nesse sentido, temos que as pescas realizadas em locais mais distantes da linha do equador e, consequentemente, mais próximos dos polos da terra obtiveram menores proporções de atum tropical dentre os peixes capturados. Análise semelhante apontou não haver indícios de correlação entre a longitude e a variável resposta, uma vez que, a 95% de confiança, a estimativa intervalar para o coeficiente de correlação de Pearson foi $[-0{,}16; 0{,}17]$.

As temperaturas observadas variam entre $11{,}58^\circ$ e $27{,}61^\circ$, com mediana de $24{,}12^\circ$. O intervalo de 95% de confiança obtido para o coeficiente de correlação de Pearson foi $[0{,}38; 0{,}62]$. Com p-valor menor que 0,001, o teste de hipóteses apontou evidências de correlação positiva moderada ente a temperatura e variável resposta. Assim, há evidências que a proporção de atum tropical capturado aumenta em locais onde a temperatura da superfície do mar é maior. Observou-se também evidências de forte correlação negativa $[-0{,}86; -0{,}75]$ entre a temperatura e a latitude absoluta, no sentido de que quanto maior

for a latitude absoluta, menor é a temperatura da superfície do mar. Esses resultados também podem ser visualizados por meio da Figura 17, na qual apresentamos os gráficos de dispersão entre a temperatura e a proporção de atum tropical, e entre a temperatura e a latitude absoluta.

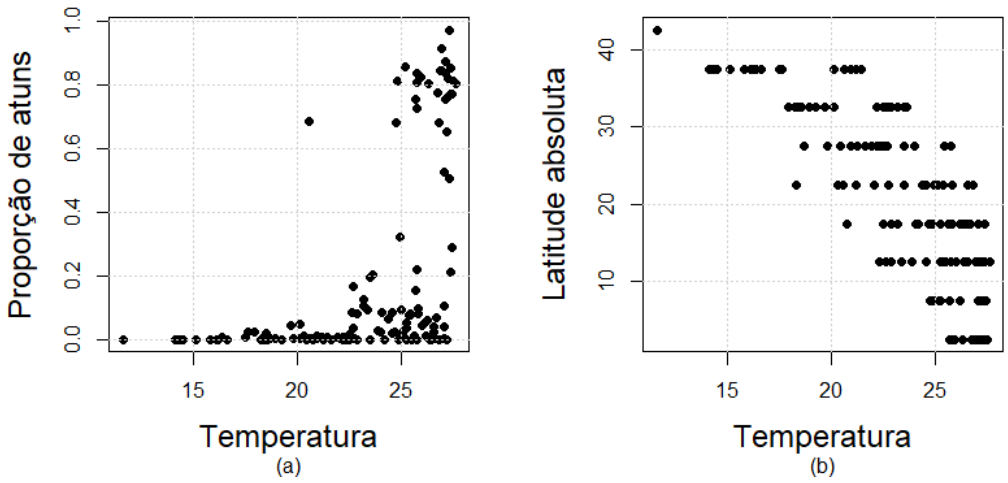

Figura 17 Diagramas de dispersão entre a temperatura e a variável resposta (a) e entre a temperatura e a latitude absoluta (b).

Apesar da forte correlação entre temperatura e latitude absoluta, inicialmente optou-se por ajustar um modelo de regressão BEZI utilizando a temperatura e a latitude absoluta como covariáveis. Admitindo-se que as realizações $y_1, \ldots, y_n$ da resposta (proporção de atum tropical capturado) são variáveis aleatórias independentes tal que cada y_t, $t = 1, \ldots, n$, tem distribuição BEZI na forma expressa em (7.2.4), com parâmetros μ_t, σ_t e ν_t, foram consideradas as estruturas de regressão

$$\begin{aligned} g_\mu(\mu_t) &= \beta_1 + \beta_2 x_{t1} + \beta_3 x_{t2}, \\ g_\sigma(\sigma_t) &= \gamma_1 + \gamma_2 x_{t1} + \gamma_3 x_{t2}, \\ g_\nu(\nu_t) &= \rho_1 + \rho_2 x_{t1} + \rho_3 x_{t2}, \end{aligned} \tag{8.2.5}$$

em que β, γ e ρ são vetores contendo os parâmetros de regressão dos três submodelos, e x_{t1} e x_{t2} são, respectivamente, os valores da temperatura e da latitude absoluta assumidos pela t-ésima observação. Para fins de facilitar a interpretação, foram utilizadas como ligação a função logit para $g_\mu(\cdot)$ e $g_\sigma(\cdot)$, e a função logarítmica para $g_\nu(\cdot)$. Esse modelo será denominado de Modelo 1.

Considerando a forte correlação entre as covariáveis, também foram ajustados outros três modelos encaixados ao Modelo 1, para fins de comparação da adequabilidade do ajuste. O Modelo 2 foi ajustado utilizando somente a covariável temperatura nas três

estrutura de regressão, ou seja, considerando que $\beta_3 = \gamma_3 = \rho_3 = 0$, e o Modelo 3 somente a covariável latitude absoluta ($\beta_2 = \gamma_2 = \rho_2 = 0$). Ainda, foi ajustado um quarto modelo, análogo ao Modelo 1, porém utilizando somente a covariável temperatura na estrutura de regressão associada a ν_t, ou seja, considerando $\rho_3 = 0$.

A Figura 18 apresenta os *worm plots* dos 4 modelos ajustados. Conforme se observa, os Modelos 1 e 4 foram os que apresentaram os melhores ajustes, uma vez que para ambos os pontos estão todos dentro dos limites das bandas de confiança e seguem a linha horizontal ao centro, sem comportamento sistemático evidente. O Modelo 2 não apresentou grandes indícios de afastamento da suposição da distribuição da variável resposta, uma vez que os pontos também estão dentro dos limites das bandas, entretanto, percebe-se um comportamento sistemático. O Modelo 3 foi o que apresentou os maiores desvios em relação às suposições, uma vez que apresentou pontos fora dos limites das bandas.

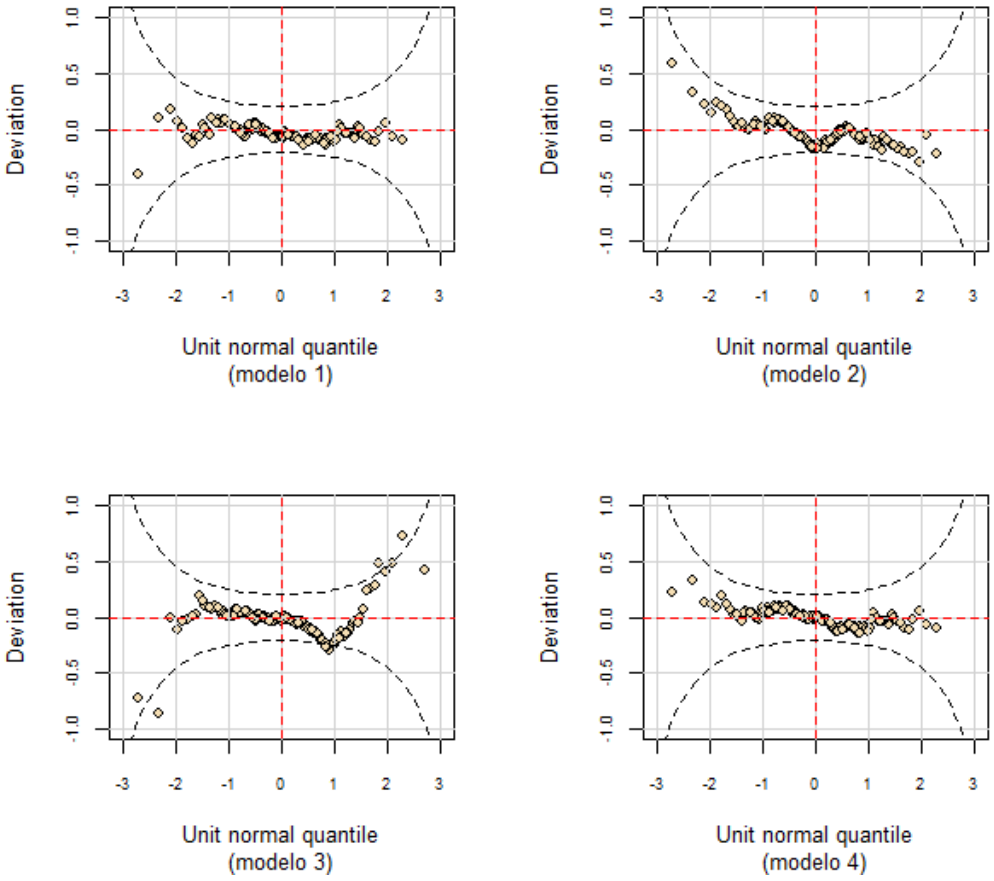

Figura 18 *worm plots* dos modelos de regressão BEZI ajustados.

Foram efetuados testes da razão de verossimilhanças utilizando o Modelo 1 como

o modelo de referência. A Tabela 9 elenca os valores observados das estatísticas de teste, graus de liberdade e respectivos p-valores. Conforme os resultados observados, existem evidências de melhor ajuste do Modelo 1 em comparação aos Modelos 2 e 3, conforme foi percebido por meio dos *worm plots*. Entretanto, o teste não apontou evidências suficientes para rejeição do Modelo 4 em favor do Modelo 1.

Tabela 9: Resultados dos testes da razão de verossimilhanças para comparação entre o Modelo 1, e os demais modelos ajustados.

Modelo sob H_0	Modelo sob H_1	Est. Teste	Graus de Liberdade	p-valor
2	1	28,62	3	$< 0{,}001$
3	1	15,666	3	0,001
4	1	0,03	1	0,874

Uma vez que os Modelos 1 e 4 apresentaram, até aqui, indícios de adequabilidade de seus ajustes, então optamos por seguir a análise utilizando o Modelo 4, que, além de ter menor complexidade, apresentou os valores ligeiramente menores para os critérios de informação. Enquanto o Modelo 1 apresentou AIC de $-98{,}38$, BIC de $-71{,}91$ e CAIC de $-85{,}14$, o Modelo 4 apresentou, respectivamente, $-100{,}35$, $-76{,}82$ e $-88{,}59$ para os mesmos critérios. A Tabela 10 apresenta as estimativas para os parâmetros do modelo selecionado, além dos erros padrão, estatısticas z e p-valores do teste de Wald para a nulidade dos coeficientes da regressão. Observa-se que os coeficientes das covariáveis em todos os submodelos apresentaram significância estatística a um nível de 5%.

Tabela 10: Estimativas dos coeficientes de regressão dos modelos ajustados, além dos erros padrão, estatısticas z e p-valores do teste de Wald para a nulidade dos coeficientes da regressão.

Parâmetro	Estimativa	Erro padrão	Est. z	p-valor
β_1	-4,75	1,67	-2,84	0,005
β_2	0,19	0,06	3,25	0,001
β_3	-0,09	0,02	-5,31	$< 0{,}001$
γ_1	-1,68	1,24	-1,35	0,178
γ_2	0,10	0,04	2,20	0,030
γ_3	-0,05	0,01	-3,63	$< 0{,}001$
ρ_1	3,54	1,22	2,90	0,004
ρ_2	-0,21	0,05	-3,77	$< 0{,}001$

Portanto, com base na análise efetuada, entende-se que o modelo selecionado é adequado para explicar o comportamento e variação da proporção de atum tropical capturado no Oceano Atlântico no ano de 1972 por meio da temperatura da superfície do mar e da latitude absoluta, conforme se pretendia. Assim, para as ligações escolhidas,

temos que

$$\widehat{\mu}_t = \frac{\exp(\widehat{\beta}_1 + \widehat{\beta}_2 x_{t1} + \widehat{\beta}_3 x_{t2})}{1 + \exp(\widehat{\beta}_1 + \widehat{\beta}_2 x_{t1} + \widehat{\beta}_3 x_{2t})} = \frac{\exp(-4{,}75 + 0{,}19 x_{t1} - 0{,}09 x_{t2})}{1 + \exp(-4{,}75 + 0{,}19 x_{t1} - 0{,}09 x_{t2})},$$

$$\widehat{\sigma}_t = \frac{\exp(\widehat{\gamma}_1 + \widehat{\gamma}_2 x_{t1} + \widehat{\gamma}_3 x_{t2})}{1 + \exp(\widehat{\gamma}_1 + \widehat{\gamma}_2 x_{t1} + \widehat{\gamma}_3 x_{t2})} = \frac{\exp(-1{,}68 + 0{,}10 x_1 - 0{,}05 x_{t2})}{1 + \exp(-1{,}68 + 0{,}10 x_1 - 0{,}05 x_{t2})}, \tag{8.2.6}$$

$$\widehat{\nu}_t = \exp(\widehat{\rho}_1 + \widehat{\rho}_2 x_{t1}) = \exp(3{,}54 - 0{,}21 x_{t1}).$$

De (2.2.4) e considerando a reparametrização descrita em (7.2.4), temos que a média da variável resposta é dada por

$$\mathrm{E}(y_t) = \left(1 - \frac{\nu_t}{1 + \nu_t}\right) \mu_t = \left(\frac{1}{1 + \nu_t}\right) \mu_t,$$

em que μ_t e ν_t são estimados conforme expressões em (8.2.6).

Observa-se que a relação entre a temperatura e a resposta média é positiva, no sentido de que a proporção média de atum fisgado tende a aumentar em localidades cuja temperatura da superfície do oceano é maior. Por outro lado, a relação entre a latitude absoluta e a resposta é negativa, uma vez que a proporção média de atum tende a diminuir quanto maior for a latitude.

A Tabela 11 elenca, para diferentes cenários segundo o modelo final ajustado, o impacto percentual na média da variável resposta ao acrescermos 1 unidade em uma das covariáveis mantendo a outra constante. Conforme se observa, mantendo a latitude absoluta fixada em, por exemplo, 20°, a cada 1° Celsius acrescido na temperatura, a proporção média de atum tropical capturado aumenta em aproximadamente 23,66%. Esse impacto tende a ser suavemente maior quanto maior for o valor fixado para a latitude absoluta. Por outro lado, fixada a temperatura em, por exemplo, 20° Celsius, a cada unidade acrescida na latitude absoluta, a proporção média de atum tropical capturado é reduzida em 7,68%. Para esse caso, o impacto tende a ser levemente menor em temperaturas mais altas.

Ainda, de (8.2.6), temos que

$$\log\left(\frac{\widehat{\mu}_t}{1 - \widehat{\mu}_t}\right) = \widehat{\beta}_1 + \widehat{\beta}_2 x_{t1} + \widehat{\beta}_3 x_{t2} = -4{,}75 + 0{,}19 x_{t1} - 0{,}09 x_{t2},$$

e, portanto

$$\frac{\widehat{\mu}_t}{1 - \widehat{\mu}_t} = \exp(\widehat{\beta}_1 + \widehat{\beta}_2 x_{t1} + \widehat{\beta}_3 x_{t2}) = \exp(-4{,}75 + 0{,}19 x_{t1} - 0{,}09 x_{t2}).$$

Tabela 11: Impactos na média estimada da variável resposta ao acrescer 1 unidade em uma das covariáveis mantendo a outra constante, para diferentes valores fixados.

Covariável fixada	Valor fixado	Covariável acrescida	Acréscimo	Alt. resposta média (%)
Latitude absoluta	2,5°	Temperatura	23° para 24° C	17,48
Latitude absoluta	10°	Temperatura	23° para 24° C	20,66
Latitude absoluta	20°	Temperatura	23° para 24° C	23,66
Latitude absoluta	30°	Temperatura	23° para 24° C	25,31
Latitude absoluta	40°	Temperatura	23° para 24° C	26,10
Temperatura	12° C	Latitude absoluta	20° para 21°	-8,08
Temperatura	16° C	Latitude absoluta	20° para 21°	-7,95
Temperatura	20° C	Latitude absoluta	20° para 21°	-7,68
Temperatura	23° C	Latitude absoluta	20° para 21°	-7,31
Temperatura	27° C	Latitude absoluta	20° para 21°	-6,50

Além disso, observa-se que a razão $\widehat{\mu}_t/(1-\widehat{\mu}_t)$ caracteriza a chance ou *odds* estimada. Supondo que o valor da i-ésima covariável é acrescida de c unidades enquanto a outra covariável é mantida constante, então a razão de chances estimada é dada por

$$\frac{\widehat{\mu}_t^*/(1-\widehat{\mu}_t^*)}{\widehat{\mu}_t/(1-\widehat{\mu}_t)} = e^{c\widehat{\beta}_i},$$

em que $\widehat{\mu}_t^*$ é a estimativa de μ_t para o valor da covariável acrescida de c unidades, e $\widehat{\mu}_t$ é a estimativa de μ_t sob o valor anterior, ou seja, antes de acrescer c unidades. Assim, esta expressão nos permite estimar, por exemplo, a chance de captura de atum tropical em comparação às demais espécies de peixes, dados os valores para as covariáveis. Por exemplo, mantida a latitude absoluta fixa, um acréscimo em 1° Celsius na temperatura, aumenta a chance estimada de captura de atum tropical em aproximadamente 21,2%, uma vez que a razão de chances estimada é $\exp(0{,}19) = 1{,}212$. Por outro lado, mantida a temperatura fixada, a chance estimada de captura de atum tropical diminui em aproximadamente 8,2% ao acrescentar 1° à latitude absoluta, uma vez que a razão de chances estimada é $\exp(-0{,}09) = 0{,}918$.

Para o submodelo de regressão associado ao componente discreto ν_t, podemos fazer uma interpretação análoga à efetuada para o submodelo da média. Observe que conforme (7.2.1), $\nu_t = \alpha_t/(1-\alpha_t)$, em que α_t representa a probabilidade de se observar o valor 0 para a t-ésima observação. Assim, de (8.2.6) temos

$$\log(\widehat{\nu}_t) = \log\left(\frac{\widehat{\alpha}_t}{1-\widehat{\alpha}_t}\right) = \widehat{\rho}_1 + \widehat{\rho}_2 x_{t1} = 3{,}54 - 0{,}21 x_{t1},$$

que equivale a

$$\widehat{\nu}_t = \frac{\widehat{\alpha}_t}{1-\widehat{\alpha}_t} = \exp(\widehat{\rho}_1 + \widehat{\rho}_2 x_{t1}) = \exp(3{,}54 - 0{,}21 x_{t1}).$$

Supondo que a temperatura (x_{t_1}) é acrescida de d unidades, então a razão de chances estimada para esse submodelo é dada por

$$\frac{\widehat{\nu}_t^*}{\widehat{\nu}_t} = \frac{\widehat{\alpha}_t^*/(1-\widehat{\alpha}_t^*)}{\widehat{\alpha}_t/(1-\widehat{\alpha}_t)} = e^{d\widehat{\rho}_2},$$

em que $\widehat{\nu}_t^*$ é a estimativa de ν_t para o valor da covariável acrescida de d unidades, e $\widehat{\nu}_t$ é a estimativa de ν_t sob o valor anterior. Dessa expressão temos que, por exemplo, um acréscimo em 1° Celsius na temperatura, reduz a chance estimada de não termos nenhum atum tropical dentre os capturados em 18,9%, uma vez que a razão de chances estimada é $\exp(-0{,}21) = 0{,}811$.

Por fim, a título comparativo, ajustamos um modelo de regressão beta com precisão variável e sem inflação. Para isso foi utilizado o mesmo conjunto de dados desta aplicação, porém, com os valores da variável resposta que assumem 0 alterados para 0,001, de modo que a variável resposta ficou restrita ao intervalo (0,1). Desse modo, considerando a parametrização expressa em (7.2.3) as estruturas de regressão são

$$g_\mu(\mu_t) = \beta_1 + \beta_2 x_{t1} + \beta_3 x_{t2},$$
$$g_\sigma(\sigma_t) = \gamma_1 + \gamma_2 x_{t1} + \gamma_3 x_{t2}.$$

A Figura 19 apresenta o *worm plot* dos resíduos quantílicos do modelo de regressão beta sem inflação ajustado após a alteração nos valores da variável resposta. Observa-se que o modelo não ficou bem ajustado, uma vez que é possível identificar pontos fora dos limites das bandas de confiança, e um comportamento sistemático bem evidente. Desse modo, temos indícios de afastamentos graves da suposição da distribuição da variável resposta. Além disso, a forma da curva gerada pelos pontos permite algumas interpretações adicionais. A primeira é que, como a curva apresenta variações bruscas de posição em relação ao eixo vertical e não segue próxima à linha horizontal ao longo da origem, temos que em alguns momentos a média pode estar sendo subestimada, enquanto em outros esta pode estar consideravelmente superestimada. Observamos também que até a linha vertical na origem, a variância pode estar sendo superestimada e, após esse ponto, subestimada. Além disso, a curva forma claramente uma parábola (forma de 'U'), o que indica um excesso de assimetria à esquerda em relação ao esperado para um modelo bem ajustado.

Logo, para as condições dos dados utilizados nesta aplicação, verifica-se que a alteração das observações cuja variável resposta assume valor zero para viabilizar a utilização do modelo de regressão beta convencional (sem inflação) se mostra uma abordagem inadequada. Nesse caso, foram evidenciados alguns problemas que podem surgir ao utilizar tal artifício, ficando ilustrado que os melhores resultados são obtidos ao usar os dados

em sua forma original, e ajustando um modelo de regressão adequado.

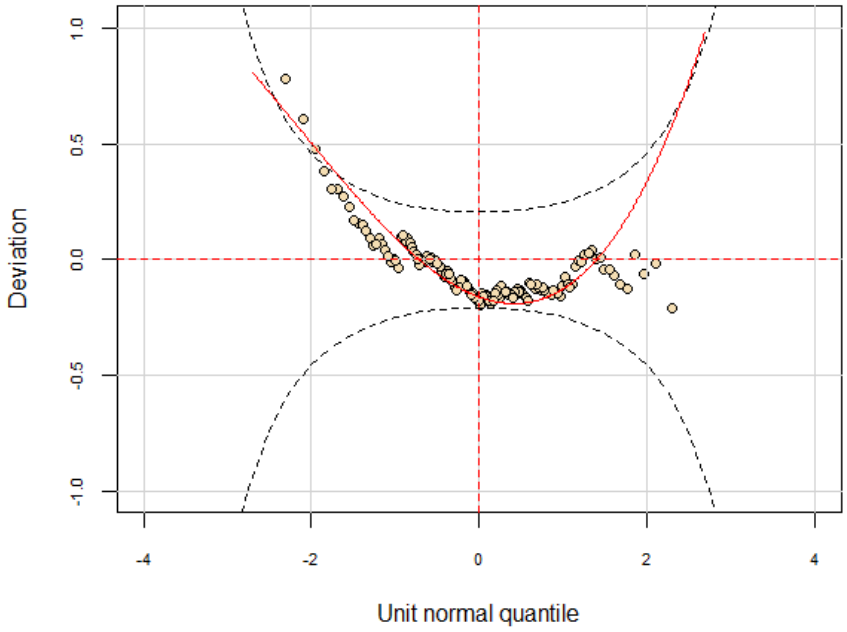

Figura 19 *worm plot* do modelo de regressão beta sem inflação ajustado.

9 Considerações finais

Com o objetivo de estudar técnicas adequadas para modelagem de dados contínuos limitados ao intervalo unitário, foram apresentados neste trabalho o modelo de regressão beta linear com precisão constante introduzido por Ferrari e Cribari-Neto (2004), o modelo de regressão beta linear com precisão variável de Smithson e Verkuilen (2006), e os modelos de regressão beta inflacionados lineares desenvolvidos por Ospina e Ferrari (2012). Estes modelos de regressão foram comparados por meio de suas principais características e propriedades teóricas para a correta identificação das principais aplicações, vantagens e limitações na sua utilização para modelagem de dados limitados ao intervalo unitário. Dados que representam taxas e proporções são exemplos de aplicações comuns para tais modelo de regressão.

Primeiro, foi identificado que para a modelagem de dados contínuos limitados ao intervalo (0,1), o modelo de regressão beta com precisão constante (FERRARI; CRIBARI-NETO, 2004) e o modelo de regressão beta com precisão variável (SMITHSON; VERKUILEN, 2006) são abordagens que podem ser adequadas. Além disso, vimos que a modelagem do parâmetro de precisão representou uma melhoria na técnica, uma vez que, a despeito da maior complexidade, tende a deixar o modelo mais flexível e, consequentemente, com maior possibilidade de obtenção de um bom ajuste.

Na sequência, vimos que os modelos de regressão beta inflacionados (OSPINA; FERRARI, 2012), desenvolvidos com o objetivo de acomodar os valores 0 e/ou 1 à distribuição beta, surgiram a partir de uma limitação existente nos modelos citados anteriormente, uma vez que o suporte da distribuição beta não contempla tais valores. Estes modelos foram aplicados a dados reais e simulados utilizando as implementações existentes no software `R`. Nos estudos de simulação foi identificado que a utilização do modelo de regressão beta com precisão constante para dados que possuam a característica de precisão variável, pode introduzir vieses significativos nas estimativas dos coeficientes de regressão, e, desse modo, prejudicar a correta interpretação da relação entre a variável resposta e as covariáveis. Não obstante, as predições para os dois modelos apresentaram resultados próximos. Desse modo, caso o objetivo do estudo seja, por exemplo, somente obter a predição, sem interesse pela interpretabilidade da relação entre as variáveis, então o modelo de regressão beta com precisão constante pode ser uma alternativa a ser avaliada.

Por fim, foram efetuadas duas aplicações a dados reais. Na primeira, observou-se que o modelo de regressão beta sem inflação com precisão constante é adequado para modelar a relação entre consumo máximo de energia dos eletrodomésticos, em megawatts, e a resposta fator de simultaneidade para sistemas prediais de gás natural, que é limitada a (0,1). Também, vimos que o modelo de regressão linear normal não se mostrou adequado,

uma vez que ajustou valores negativos para o fator de simultaneidade. Além disso, o ajuste de um modelo de regressão linear com a transformação logit na variável resposta produziu um ajuste razoável, entretanto, impossibilita a interpretabilidade direta em termos da média da variável resposta e coeficiente estimado da covariável.

Na segunda aplicação, onde a variável resposta assumiu alguns valores iguais a zero, verificamos que o modelo ajustado que apresentou os melhores resultados foi a regressão BEZI com precisão variável. Assim, por meio deste modelo de regressão, obtivemos que a proporção de atum tropical capturado tende a aumentar quando a temperatura do oceano é maior. Em contrapartida, a proporção de atum tropical fisgado no palangre tende a reduzir quanto maior for a latitude absoluta do local da pesca. Além disso, obtivemos que a chance estimada de não termos nenhum atum tropical dentre os capturados reduz consideravelmente a cada 1^{o} Celsius acrescidos na temperatura da superfície do mar. Finalmente, identificamos que a alteração de observações cuja variável resposta assumia o valor zero para viabilizar a utilização da regressão beta convencional (sem inflação) se mostrou uma abordagem inadequada, uma vez que, prejudicou a qualidade do ajuste geral.

Referências

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974.

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

ATKINSON, A. C. *Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis*. [S.l.], 1985.

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

BAYER, F. M. Modelagem e inferência em regressão beta. Universidade Federal de Pernambuco, 2011.

BICKEL, P. J.; DOKSUM, K. A. *Mathematical Statistics: Basic ideas and Selected Topics*. [S.l.]: Prentice Hall, 2001.

BLOM, G. Statistical estimates and transformed beta-variables wiley. *Almquist und Wiksell, New York/Stockholm*, 1958.

BUSE, A. The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, Taylor & Francis, v. 36, n. 3a, p. 153–157, 1982.

BUUREN, S. v. Worm plot to diagnose fit in quantile regression. *Statistical Modelling*, Sage Publications India Pvt. Ltd, B-42, Panchsheel Enclave, New Delhi, v. 7, n. 4, p. 363–376, 2007.

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

CASELLA, G.; BERGER, R. L. *Inferência Estatística*. [S.l.]: CENGAGE Learning, 2011. 95–96;421–422 p.

CHARNET, R. et al. *Análise de modelos de regressão linear com aplicações*. [S.l.: s.n.], 1999.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Piracicaba: USP*, 2008.

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.

CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in r. *Journal of statistical software*, v. 34, p. 1–24, 2010.

DRAPER, N. R.; SMITH, H. *Applied Regression Analysis*. [S.l.]: John Wiley & Sons, 1998. v. 326.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and graphical statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.

ESPESCA. La pesca al palangre. 2023. Disponível em: ⟨https://espesca.com/palangre/⟩.

ESPINHEIRA, P. L. *Regressão beta*. Tese (Doutorado) — Universidade de São Paulo, 2007.

FERNANDES, V. V. Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão. Universidade Federal de São Carlos, 2019.

FERRARI, S. L. Beta regression. *Wiley StatsRef: Statistics Reference Online*, Wiley Online Library, p. 1–5, 2014.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

GUMBEL, E. J. *Statistical theory of extreme values and some practical applications: a series of lectures*. [S.l.]: US Government Printing Office, 1954. v. 33.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models london chapman and hall. *Inc*, 1990.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. *Continuous univariate distributions, volume 2*. [S.l.]: John wiley & sons, 1995. v. 289.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling*, Sage Publications Sage CA: Thousand Oaks, CA, v. 3, n. 3, p. 193–213, 2003.

KODA, C. A. Modelos mistos para respostas positivas aumentadas em zero. 2018.

MACYS, J. Stability of characterization of a degenerate distribution. *Lithuanian Mathematical Journal*, Springer, v. 27, n. 1, p. 76–82, 1987.

MONLLOR-HURTADO, A.; PENNINO, M. G.; SANCHEZ-LIZASO, J. L. Shift in tuna catches due to ocean warming. *PloS one*, Public Library of Science San Francisco, CA USA, v. 12, n. 6, p. e0178196, 2017.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. [S.l.]: Saraiva Educação SA, 2017.

NELDER, J. A.; LEE, Y. Generalized linear models for the analysis of taguchi-type experiments. *Applied stochastic models and data analysis*, Wiley Online Library, v. 7, n. 1, p. 107–120, 1991.

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

OSPINA, R. *Estimação pontual e intervalar em um modelo de regressão beta*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2004.

OSPINA, R. *Modelos de regressão beta inflacionados*. Tese (Doutorado) — Universidade de São Paulo, 2008.

OSPINA, R.; FERRARI, S. L. Inflated beta distributions. *Statistical Papers*, Springer, v. 51, n. 1, p. 111, 2010.

OSPINA, R.; FERRARI, S. L. P. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 6, p. 1609–1623, 2012.

PEREIRA, T. L. Regressão beta inflacionada: Inferência e aplicações. Universidade Federal de Pernambuco, 2010.

QUEIROZ, F. F. d. *Análise de dados com suporte limitado: modelos power logit e contribuições à inferência robusta*. Tese (Doutorado) — Universidade de São Paulo, 2022.

R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna*. 2022.

RIBEIRO, T. K. A.; FERRARI, S. L. P. Robust estimation in beta regression via maximum l_q-likelihood. *Statistical Papers*, Springer, p. 1–33, 2022.

ROSS, S. *Probabilidade: um curso moderno com aplicações*. [S.l.]: Bookman Editora, 2009.

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

SILVA, A. R. d. S. Modelos de regressão beta retangular heteroscedásticos aumentados em zeros e uns. [sn], 1989.

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. 54–71, 2006.

SMYTH, G. K.; VERBYLA, A. P. Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics: The official journal of the International Environmetrics Society*, Wiley Online Library, v. 10, n. 6, p. 695–709, 1999.

SONG, P. X.-K.; TAN, M. Marginal models for longitudinal continuous proportional data. *Biometrics*, Wiley Online Library, v. 56, n. 2, p. 496–502, 2000.

STASINOPOULOS, M.; RIGBY, B.; STASINOPOULOS, M. M. The gamlss package. *R help files*, 2006.

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

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, JSTOR, v. 54, n. 3, p. 426–482, 1943.

WIKIPEDIA. *Wikipedia*. [S.l.]: PediaPress, 2023.