Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

André Bisca Ferreira

A Distribuição New Unit-Lindley Inflacionada em Zero, em Um e em Zero e Um

Maringá – PR
2020

André Bisca Ferreira

A Distribuição New Unit-Lindley Inflacionada em Zero, em Um e em Zero e Um

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do centro de ciências exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de mestre em Bioestatística.

Orientador: Prof. Dr. Josmar Mazucheli

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Maringá – PR

2020

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Ferreira, André Bisca

F383d A distribuição *New Unit-Lindley* inflacionada Em Zero, Em Um e Em Zero e Um / André Bisca Ferreira. -- Maringá, PR, 2021.

49 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Josmar Mazucheli.

Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2021.

1. Modelos inflacionados (Estatística). 2. Estimação via máxima verossimilhança (Estatística). 3. Simulação Monte Carlo. 4. Modelo de regressão (Estatística). I. Mazucheli, Josmar, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 23.ed. 519.54

Márcia Regina Paiva de Brito - CRB-9/1267

ANDRÉ BISCA FERREIRA

A Distribuição *New Unit-Lindley* Inflacionada em Zero, em Um e em Zero e Um

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Prof. Dr. Josmar Mazucheli
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Wesley Bertoli da Silva
Universidade Tecnológica Federal do Paraná – UTFPR/Curitiba-PR

Prof. Dr. Diogo Francisco Rossoni
Universidade Estadual de Maringá – PBE/UEM

Maringá, 17 de agosto de 2021.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Lista de ilustrações

Figura 2.1 – F.d.p. NUL e UL assumindo diferentes valores para média μ. 13
Figura 3.1 – Comportamento da densidade NULIZ assumindo diferentes valores para o vetor paramétrico (μ, σ). (linha pontinhada em preto: proporção de zeros; linha pontilhada em azul: média da distribuição New unit-Lindley; linha pontilhada em vermelho: média conjunta do modelo de mistura) 16
Figura 3.2 – Comportamento da densidade NULIU assumindo diferentes valores para o vetor paramétrico (μ, σ). (linha pontinhada em preto: proporção de uns; linha pontilhada em azul: média da distribuição New unit-Lindley; linha pontilhada em vermelho: média conjunta do modelo de mistura) 17
Figura 6.1 – Comportamento da densidade NULIZU assumindo diferentes valores para o vetor paramétrico (μ, σ, ρ).(linha pontinhada em preto: proporção de zeros ou uns; linha pontilhada em azul: média da distribuição New unit-Lindley; linha pontilhada em vermelho: média conjunta do modelo de mistura) . . 23
Figura 9.1 – Viés das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZ (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2). 28
Figura 9.2 – EQM das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZ (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2). 28
Figura 9.3 – Viés das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZ mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$). 29
Figura 9.4 – EQM das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZ mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$). 29

Figura 9.5 – Viés das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2). 29
Figura 9.6 – EQM das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2). 30
Figura 9.7 – Viés das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIU mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$). 30
Figura 9.8 – EQM das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIU mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$). 30
Figura 9.9 – Viés das estimativas $\widehat{\beta}$, $\widehat{\gamma}$ e $\widehat{\lambda}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2). 31
Figura 9.10–EQM das estimativas $\widehat{\beta}$, $\widehat{\gamma}$ e $\widehat{\lambda}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2). 31
Figura 9.11–Viés das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZU mantendo σ e ρ fixos (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.4$, linha pontilhada: $\sigma = 0.6$ e linha tracejada-pontilhada: $\sigma = 0.8$). . 32
Figura 9.12–EQM das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZU mantendo σ e ρ fixos (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.4$, linha pontilhada: $\sigma = 0.6$ e linha tracejada-pontilhada: $\sigma = 0.8$). . 33
Figura 10.1–*Worm plot* dos resíduos para modelos MRNULIZ, MRULIZ e MRBEIZ. . . 35

Lista de abreviaturas e siglas

EMV	ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA
EQM	ERRO QUADRÁTICO MÉDIO
F.D.A.	FUNÇÃO DE DISTRIBUIÇÃO ACUMULADA
F.D.P.	FUNÇÃO DENSIDADE DE PROBABILIDADE
MRBEIZ	MODELO DE REGRESSÃO BETA INFLACIONADO EM ZERO
MRNULIU	MODELO DE REGRESSÃO NEW UNIT-LINDLEY INFLACIONADO EM UM
MRNULIZ	MODELO DE REGRESSÃO NEW UNIT-LINDLEY INFLACIONADO EM ZERO
MRNULIZU	MODELO DE REGRESSÃO NEW UNIT-LINDLEY INFLACIONADO EM ZERO E UM
NUL	DISTRIBUIÇÃO NEW UNIT-LINDLEY
NULIU	DISTRIBUIÇÃO NEW UNIT-LINDLEY INFLACIONADA EM UM
NULIZ	DISTRIBUIÇÃO NEW UNIT-LINDLEY INFLACIONADA EM ZERO
NULIZU	DISTRIBUIÇÃO NEW UNIT-LINDLEY INFLACIONADA EM ZERO E UM
UL	DISTRIBUIÇÃO UNIT-LINDLEY

Sumário

1 Introdução 11
2 A distribuição New Unit-Lindley 12
3 Agregando Zeros ou Uns à distribuição New Unit Lindley 13
4 Aspectos Inferências das distribuições New unit-Lindley inflacionada em Zero ou em Um 18
5 Modelo de Regressão New unit-Lindley inflacionado em Zero ou em Um 20
6 Agregando Zeros e Uns à distribuição New unit-Lindley 21
7 Aspectos Inferências da distribuição New unit-Lindley inflacionada em Zero e Um 24
8 Modelo de Regressão New unit-Lindley inflacionada em Zero e Um 26
9 Estudo de Simulação 26
10 Aplicação a Dados Reais 33
11 Conclusões 37

Referências 38

A distribuição New Unit-Lindley inflacionada em Zero, em Um e em Zero e UM

Resumo

Em diversas áreas do conhecimento, dados na forma de taxas, razões ou proporções, são medidos ou observados de maneira contínua no intervalo unitário na presença de zeros e/ou uns. Em tais casos, em razão da natureza contínua e restrita dessas variáveis, a análise paramétrica demanda distribuições de probabilidade que tenham as mesmas características. Há alguns anos, Ospina e Ferrari propuseram versões inflacionadas da distribuição Beta que viabilizaram seu uso na presença de zeros e/ou uns. Em decorrência deste trabalho outras distribuições unitárias (ou não) ganharam versões inflacionadas que estendiam sua utilização para um intervalo mais amplo. Nesse contexto, tomamos a distribuição continua New Unit-Lindley que tem seu suporte restrito ao intervalo aberto $(0, 1)$ e propomos neste trabalho suas respectivas versões inflacionadas em zero, em um e em zero e um segundo a metodologia empenhada por Ospina. Aspectos inferenciais e uma estrutura de regressão sobre estas distribuições são tratados neste trabalho bem como um estudo de Simulação Monte Carlo para avaliar a performance dos coeficientes regressores. Também investigamos algumas de suas propriedades e realizamos ao final uma aplicação a dados reais sobre a taxa de suicídio no ano de 2016.

Palavras-chave: Modelos Inflacionados, Estimação via máxima verossimilhança, Estudo de Simulação Monte Carlo, Modelo de Regressão.

Abstract

In several areas of knowledge, data in the form of rates, ratios or proportions are measured or observed continuously in the unit interval in the presence of zeros and/or ones. In such cases, due to the continuous and restricted nature of these variables , a parametric analysis demands probability distributions that have the same characteristics. A few years ago, Ospina and Ferrari proposed inflated versions of the Beta distribution that enabled its use in the presence of zeros and/or ones. As a result of this work, other unit distributions (or not) gained inflated versions that extended their use to a wider range. In this context, we take the New Unit-Lindley continuous distribution which has its support restricted to the open interval (0.1) and we propose in this work its versions inflated at zero, one and zero and one according to the methodology used by Ospina. Regression structures on these distributions are nested as well as a Monte Carlo Simulation study to evaluate the performance of coefficients regressors. We also investigated some of their properties and made an official application to actual data on a suicide rate for the year 2016.

Keywords: Inflated Models, Maximum Likelihood Estimation, Monte Carlo Simulation Study, Regression Models.

1 Introdução

Em diversas áreas do conhecimento, dados na forma de taxas, razões ou proporções são medidos ou observados de forma contínua dentro do intervalo unitário zero a um. Nestes casos, dada a natureza continua e restrita dos dados, a análise paramétrica demanda distribuições de probabilidade que tenham a mesmas características. A distribuição beta é amplamente utilizada neste tipo de análise em razão das diferentes formas que sua função de densidade pode assumir. Por outro lado, há situações reais do quotidiano onde ela e outras distribuições unitárias não se adequam bem aos dados o que leva a procura de novos modelos.

Dentre os modelos unitários propostos recentemente, temos as distribuições da família unit: unit-Weibull, unit-Gompertz e unit-Inverse Gaussian propostas a partir da transformação $Y = e^{-X}$ em que X segue, respectivamente, uma variável Weibull, Gompertz e Inverse Gaussian (MAZUCHELI; MENEZES; GHITANY, 2018; MAZUCHELI; MENEZES; DEY, 2019; GHITANY et al., 2019) e as distribuições unit-lindley e New unit-lindley propostas respectivamente a partir das transformações $Y = X\,(1+X)^{-1}$ e $Y = (1+X)^{-1}$ em que X segue uma variável Lindley (MAZUCHELI; MENEZES; CHAKRABORTY, 2019; MAZUCHELI; BAPAT; MENEZES, 2020) entre outras.

Embora as distribuições da família unit sejam soluções para diversos problemas práticos, a presença de zeros e/ou uns geralmente não são captadas por estes modelos o que consequentemente inviabiliza seu uso. Neste caso, uma alternativa viável da literatura é a adoção de modelos de mistura com as distribuições Degenerada em Zero, em Um e Bernoulli para atribuir pontos de massa nos extremos conforme a necessidade. Construídos a partir de somas convexas, estes modelos de mistura têm simultaneamente as naturezas contínua e discreta, possuem o suporte definido nos intervalos $[0,1)$, $(0,1]$ e $[0,1]$ de acordo com as distribuições adotadas na mistura e podem lidar também com problemas de excessos de zeros e/ou uns, veja (XIE; HE; GOH, 2001; OSPINA; FERRARI, 2010; CRIBARI-NETO; SANTOS, 2019; TOMARCHIO; PUNZO, 2019; LIU et al., 2020; RIVAS; CAMPOS, 2021; BAPAT; BHARDWAJ, 2021a) para mais detalhes sobre modelos de mistura.

No contexto de regressão, os modelos de natureza mista permitem avaliar a influência de covariáveis nos componentes contínuos e discreto como eventos distintos (ou separados) ou diretamente sobre a média marginal (também dita conjunta) do modelo. Ospina e Ferrari (2012) por exemplo, descreve a proporção de acidentes fatais em municípios brasileiros considerando o modelo beta inflacionada em zero em que, o componente discreto retrata a probabilidade de ocorrer (ou não) acidentes fatais, enquanto que o componente contínuo, explica a proporção de acidentes de trânsito tendo em vista que foram registradas mortes. Chai et al. (2018) apresentam uma discussão detalhada para dados de microbiologia comparando as parametrizações sobre a média marginal e média condicional (componente continuo) de um

modelo de regressão de natureza mista. Mais detalhes a respeito de estrutura de regressão sobre modelos de mistura veja (SANTOS; BOLFARINE, 2015; SANTOS; BOLFARINE, 2018; MENEZES; MAZUCHELI; BOURGUIGNON, 2021; QUEIROZ; LEMONTE, 2021; SILVA et al., 2021).

O restante deste trabalho esta organizado da seguinte forma. Na Seção 2 trazemos uma breve discussão acerca da distribuição New unit-Lindley. Nas Seções 3 e 6 apresentamos, respectivamente, suas versões com inflação unilateral (em zero ou em um) e bilateral (em zero e um). Nas Seções 4 e 7 discutimos aspectos infer encias relativos aos modelos enquanto que nas Seções 5 e 8 estudamos suas respectivas estruturas de regressão. Um estudo de simulação Monte Carlo é apresentado na Seção 9 para avaliar a performance das estimativas dos coeficientes regressões para cada modelo. Na Seção 10 trazemos uma aplicação real a dados sobre a taxa de suicídio e na Seção 11 apresentamos nossas considerações finais sobre o trabalho realizado.

2 A distribuição New Unit-Lindley

A distribuição New unit-Lindley (NUL) foi proposta por Mazucheli, Bapat e Menezes (2020) a partir da transformação $X = (1+Y)^{-1}$, em que $Y \sim$ Lindley(θ) (LINDLEY, 1958). Deste modo, como $Y \in (\,0,\,+\infty)$, a variável X fica então restrita ao intervalo aberto $(0,1)$ e suas funções de distribuição acumulada (f.d.a.) e densidade de probabilidade (f.d.p.) são dadas respectivamente por

$$\mathsf{F}(x \mid \theta) = \frac{(\theta + x)}{x\,(1+\theta)} \exp\left(-\frac{\theta\,(1-x)}{x}\right), \tag{2.1}$$

e

$$\mathsf{f}(x \mid \theta) = \frac{\theta^2}{(1+\theta)\,x^3} \exp\left(-\frac{\theta\,(1-x)}{x}\right) \tag{2.2}$$

em que $\theta > 0$. Por outro lado, como $\mathbb{E}(X) = \mu = \theta(1+\theta)^{-1}$, as Equações 2.1 e 2.2 podem ser reescritas na forma

$$\mathsf{F}(x \mid \mu) = \frac{x\,(1-\mu) + \mu}{x} \exp\left(-\frac{\mu\,(1-x)}{(1-\mu)\,x}\right), \tag{2.3}$$

e

$$\mathsf{f}(x \mid \mu) = \frac{\mu^2}{(1-\mu)\,x^3} \exp\left(-\frac{\mu\,(1-x)}{(1-\mu)\,x}\right), \tag{2.4}$$

em que $0 < \mu < 1$.

A distribuição NUL conta com diversas formas assimétricas unimodais, formas fechadas para os momentos e função quantil e também membro da família exponencial. Sua f.d.p. toma formas reflexivas à densidade unit-Lindley (UL), proposta por Mazucheli, Menezes e Chakraborty (2019), uma vez que a variável UL provém da transformação $X' = Y(1+Y)^{-1}$ em que $Y \sim$ Lindley(θ). Dessa forma, as distribuições NUL e UL estão relacionadas pela transformação $X = 1 - X'$, possibilitando então obter versões inflacionadas da distribuição NUL por meio desta mesma transformação sobre versões inflacionadas da distribuição UL e vice versa. Na Figura 2.1 apresentamos o comportamento das densidades NUL e UL para diferentes valores de μ sendo $\mu = \mathbb{E}(X) = \mathbb{E}(X')$. Observe que $fnul(x \mid \mu)$ e $ful(x \mid 1-\mu)$ possuem apresentam curvas espelhadas o que ilustra a relação reflexiva entre essas distribuições.

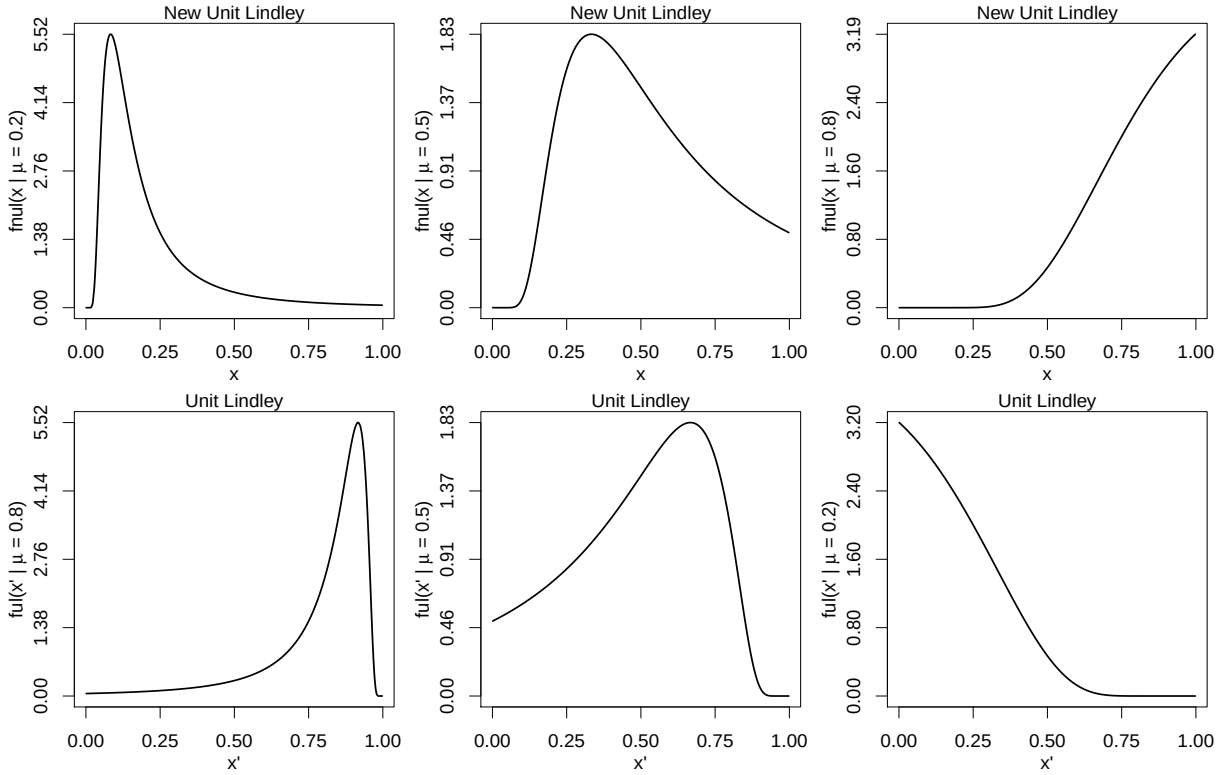

Figura 2.1 – F.d.p. NUL e UL assumindo diferentes valores para média μ.

3 Agregando Zeros ou Uns à distribuição New Unit Lindley

Embora seja possível agregar zeros e/ou uns à distribuição NUL através transformações diretas sobre as variáveis UL inflacionada em zero, em um e em zero (BAPAT; BHARDWAJ, 2021b), optamos por faze-lo via metodologia de Ospina e Ferrari (2010) que reside na construção de modelos de mistura por meio de combinações convexas com distribuições apropriadas. Por exemplo, para agregar zeros ou uns a distribuição NUL, constrói-se o modelo de mistura com uma distribuição degenerada em c, sendo c igual a zero ou a um dependendo do caso.

Deste modo, segue que as funções f.d.a. e f.d.p. desses modelos são dados respectivamente por

$$\mathsf{FNULIC}(y \mid \mu, \sigma) = \sigma\, \Delta_c(y) + (1-\sigma)\, \mathsf{F}(y \mid \mu),$$

e

$$\mathsf{fnulic}(y \mid \mu, \sigma) = \sigma\, \delta_c(y) + (1-\sigma)\, \mathsf{f}(y \mid \mu), \tag{3.1}$$

em que $y \in (0,1) \cup \{c\}$, $0 < \sigma < 1$ corresponde a massa de probabilidade em $y = c$, $\delta_c(y)$ é uma função indicadora dada por $\delta_c(y) = 1$ se $y = c$ e $\delta_c(y) = 0$ caso contrario, $\Delta_c(y)$ é a função acumulada de $\delta_c(y)$ dada por $\Delta_c(y) = 1$ se $y \geq c$ e $\Delta_c(y) = 0$ se $y < c$ e os termos $\mathsf{F}(\cdot \mid \mu)$ e $\mathsf{f}(\cdot \mid \mu)$ são, respectivamente, as funções de distribuição acumulada e densidade de probabilidade NUL, dadas pelas Equações 2.3 e 2.4.

Definição 1 *Seja Y uma variável aleatória cuja densidade é dada pela Equação 3.1, então*

- *Se $c = 0$, Y é dito ser uma distribuição New unit-Lindley inflacionada em Zero (NULIZ), e denotamos por $Y \sim NULIZ(\mu, \sigma)$ em que $\sigma = P(Y = 0)$.*
- *Se $c = 1$, Y é dito ser uma distribuição New unit-Lindley inflacionada em Um (NULIU), e denotamos por $Y \sim NULIU(\mu, \sigma)$ em que $\sigma = P(Y = 1)$.*

Observe que umas das desvantagens deste modelo de mistura, e consequentemente das distribuições NULIZ e NULIU, é a descontinuidade no ponto de massa em $x = c$. Embora o suporte[1] da distribuição NUL tenha sido ampliado em c, neste ponto, a variável tem uma natureza discreta enquanto que para o restante das observações a natureza é contínua. Sendo assim, nosso modelo tem simultaneamente uma natureza continua e discreta e por isso, quando as densidades máximas dos componentes contínuos e discretos são iguais, o modelo assume uma forma bimodal. Posto isto, temos que a distribuição NULIZ é

- bimodal com moda em $y = 0$ e $y = \mu\,(3(1-\mu))^{-1}$ se $\mu < 3/4$ e $\sigma = \mathbf{b}(\mu)/(\mathbf{b}(\mu) + \mu)$
- bimodal com moda em $y = 0$ e $y = 1$ se $\mu \geq 3/4$ e $\sigma = \mu^2\,(\mu^2 - \mu + 1)^{-1}$
- unimodal com moda em $y = 0$ se $\mu \geq 3/4$ e $\sigma > \mu^2\,(\mu^2 - \mu + 1)^{-1}$ ou se $\mu < 3/4$ e $\sigma > \mathbf{b}(\mu)/(\mathbf{b}(\mu) + \mu)$
- unimodal com moda em $y = \mu\,(3(1-\mu))^{-1}$ se $\mu < 3/4$ e $\sigma < \mathbf{b}(\mu)/(\mathbf{b}(\mu) + \mu)$

[1] Conjunto de todas as realizações da variável aleatória para as quais a função densidade de probabilidade é maior do que zero.

- unimodal com moda em $y = 1$ se $\mu \geq 3/4$ e $\sigma < \mu^2\,(\mu^2 - \mu + 1)^{-1}$

enquanto que a distribuição NULIU é

- bimodal com moda em $y = (3(1-\mu))^{-1}$ e $y = 1$ se $\mu < 3/4$ e $\sigma = \mathbf{b}(\mu)/(\mathbf{b}(\mu) + \mu)$
- unimodal com moda em $y = \mu\,(3(1-\mu))^{-1}$ se $\mu < 3/4$ e $\sigma < \mathbf{b}(\mu)/(\mathbf{b}(\mu) + \mu)$
- unimodal com moda em $y = 1$ se $\mu \geq 3/4$ ou se $\mu < 3/4$ e $\sigma > \mathbf{b}(\mu)/(\mathbf{b}(\mu) + \mu)$

em que $\mathbf{b}(\mu) = 27 \exp\left[(4\mu - 3)(1-\mu)^{-1}\right](\mu - 1)^2$.

Observe que, se Y é uma variável aleatória com f.d.p. dada pela Equação 3.1, então $\mathbb{E}[Y^r] = \mu_r^{'} = \mathbb{E}[(1-\sigma)\,X^r + \sigma\,Z^r] = (1-\sigma)\,\mathbb{E}[X^r] + \sigma\,\mathbb{E}[Z^r] = (1-\sigma)\,\mu_r + \sigma c$, em que r é um valor inteiro maior que zero; X é uma variável aleatória NUL com momento de ordem r igual μ_r e Z é uma variável aleatória degenerada em c. Em particular, para $r = 1, 2$ temos que a média e variância são dadas respectivamente por

$$\begin{aligned}
\mu_1^{'} &= (1-\sigma)\,\mu + \sigma c \\
\mu_2^{'} &= (1-\sigma)\left[\mu^2\,\mathbf{d}(\mu)(1-\mu)^{-1}\right] + \sigma c
\end{aligned}$$

em que $\mathbf{d}(\mu) = \mathrm{Ei}[1,\, \mu(1-\mu)^{-1}]\,\exp\left(\mu(1-\mu)^{-1}\right)$ e $\mathrm{Ei}[a, z] = \int_1^\infty z^{-a} e^{-yz} dz$ é a função integral exponencial (ABRAMOWITZ; STEGUN, 1974). Posto isso, dependendo do tipo de inflacionamento realizado, a média marginal do modelo, $\mathbb{E}[Y]$, desloca-se para a esquerda ou direita. Nas Figuras 3.1 e 3.2 apresentamos o comportamento gráfico das densidades NULIZ e NULIU para diferentes valores de μ e σ e, assinalamos pelas linhas pontilhadas em azul e vermelho as respectivas médias das distribuição NUL e do modelo de mistura.

Proposição 1 *As distribuições NULIZ e NULIU pertencem a família exponencial bi-paramétrica.*

Prova: De fato, sendo $\eta = (\eta_1, \eta_2)$, com $\eta_1 = \log(\sigma) - \log(1-\sigma) - \log\left(\dfrac{\mu^2}{1-\mu}\right)$ e $\eta_2 = -\dfrac{\mu}{1-\mu}$; e $T(y) = (t_1(y), t_2(y))$, com $t_1(y) = \delta_c(y)$ e $t_2(y) = \dfrac{1-y}{y}$ se $y \in (0,1)$ e 0 se $y = c$, obtém-se a Equação 3.1 escrita na forma

$$\text{fnulic}(y \mid \mu, \sigma) = \exp\{\eta^\top T(y) - \zeta(\eta)\}\, h(y)\,,$$

em que $\zeta(\eta) = \eta_1 - \log(\sigma)$ e $h(y) = y^{-3}$ se $y \in (0,1)$ e 1 se $y = c$. ■

Note que, pelas Proposições 1 e 2, as versões inflacionadas em zero, em um e em zero e um da distribuição NUL, podem ser empregadas na estrutura de regressão como um objeto da família dos Modelos Lineares Generalizados (NELDER; WEDDERBURN, 1972).

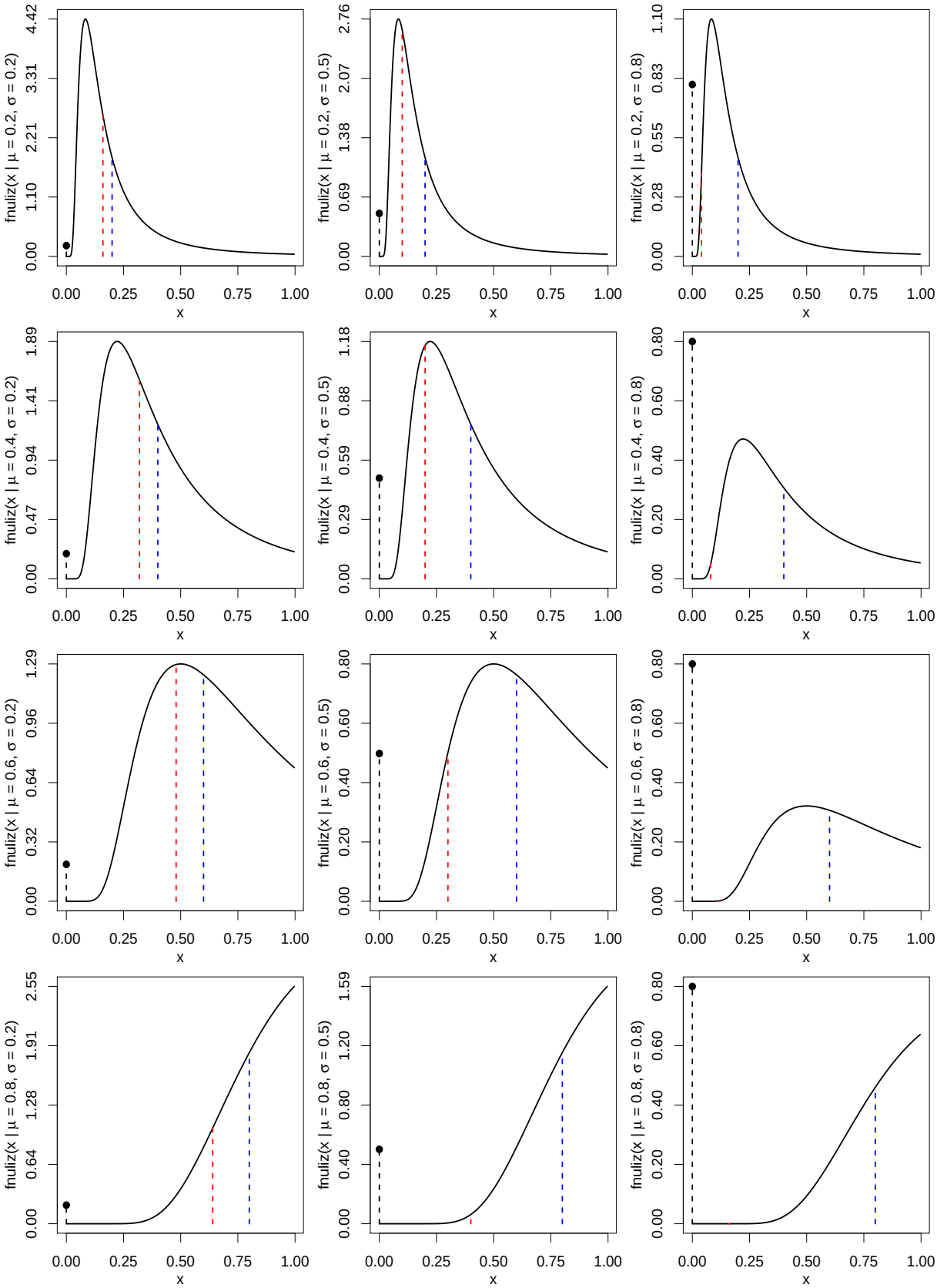

Figura 3.1 – Comportamento da densidade NULIZ assumindo diferentes valores para o vetor paramétrico (μ, σ). (linha pontinhada em preto: proporção de zeros; linha pontilhada em azul: média da distribuição New unit-Lindley; linha pontilhada em vermelho: média conjunta do modelo de mistura)

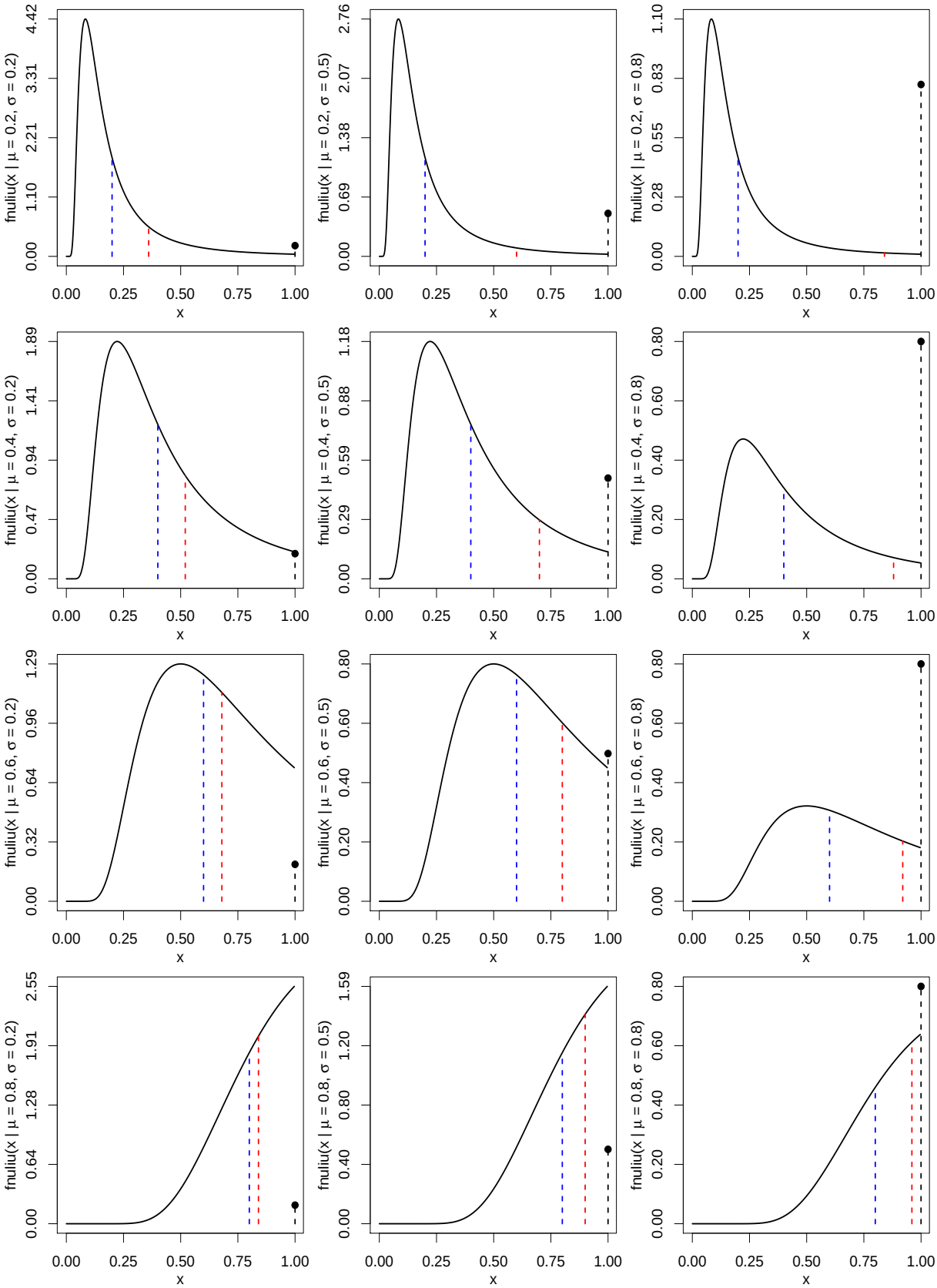

Figura 3.2 – Comportamento da densidade NULIU assumindo diferentes valores para o vetor paramétrico (μ, σ). (linha pontinhada em preto: proporção de uns; linha pontilhada em azul: média da distribuição New unit-Lindley; linha pontilhada em vermelho: média conjunta do modelo de mistura)

As distribuições NULIZ e NULIU contam com formas fechadas para os quantis sendo elas dadas respectivamente por

$$QNULIZ(p \mid \mu, \sigma) = \begin{cases} \dfrac{\mu}{\left(W_{-1}\left[-\dfrac{(p-\sigma)\,\mathbf{e}(\mu)}{(1-\sigma)}\right]+1\right)(-1+\mu)} & \text{se } \sigma < p < 1 \\ 0 & \text{se } 0 < p \leq \sigma \end{cases}$$

e

$$QNULIU(p \mid \mu, \sigma) = \begin{cases} \dfrac{\mu}{\left(W_{-1}\left[-\dfrac{p\,\mathbf{e}(\mu)}{(1-\sigma)}\right]+1\right)(-1+\mu)} & \text{se } 0 < p \leq 1-\sigma \\ 1 & \text{se } 1-\sigma < p < 1 \end{cases}$$

em que $\mathbf{e}(\mu) = (1-\mu)^{-1}\exp(-(1-\mu)^{-1})$ e W_{-1} denota o ramo negativo da função `W` de Lambert (CORLESS et al., 1996; JODRÁ, 2010).

4 Aspectos Inferências das distribuições New unit-Lindley inflacionada em Zero ou em Um

Nesta seção dispomos expressões gerais do estimador e matriz de informação esperada de Fisher para o vetor de parâmetros $\Psi = (\mu, \sigma)$ da distribuição NUL inflacionada em c. As expressões relativas às distribuições NULIZ e NULIU são obtidas tomando, respectivamente, c igual a zero e a um.

Sejam $Y = Y_1, Y_2. \ldots, Y_n$ uma amostra aleatória retirada de uma população com f.d.p. dada pela Equação 3.1. Sem perca de generalidade, a densidade da variável Y pode ser expressa na forma

$$\text{fnulic}(y \mid \mu, \sigma) = \{(1-\sigma)^{1-\delta_c(y)}\sigma^{\delta_c(y)}\}\{\text{f}(y \mid \mu)\}^{1-\delta_c(y)},$$

em que fnulic$(\cdot \mid \mu, \sigma)$, como função dos parâmetros, pode ser fatorado em dois termos independentes. Nessas condições, para $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ observado, a função de verossimilhança de $\Psi = (\mu, \sigma)$ é dada por

$$L(\Psi \mid \mathbf{y}) = L_1(\sigma \mid \mathbf{y}) \times L_2(\mu \mid \mathbf{y}), \tag{4.1}$$

em que

$$L_1(\sigma \mid \mathbf{y}) = \prod_{i=1}^{n} \sigma^{\delta_c(y_i)}(1-\sigma)^{1-\delta_c(y_i)} \tag{4.2}$$

e

$$L_2(\mu \mid \mathbf{y}) = \prod_{i=1}^{n} \left\{ \frac{\mu^2}{(1-\mu)\, y_i^3} \exp\left(-\frac{\mu\,(1-y_i)}{(1-\mu)\, y_i} \right) \right\}^{1-\delta_c(y_i)}. \tag{4.3}$$

Das Equações 4.1, 4.2 e 4.3 temos que o logaritmo da função de verossimilhança é dada por

$$\ell(\Psi \mid \mathbf{y}) = \ell_1(\sigma \mid \mathbf{y}) + \ell_2(\mu \mid \mathbf{y}),$$

sendo

$$\ell_1(\sigma \mid \mathbf{y}) = (n - \mathbf{s}(\mathbf{y})) \log(1-\sigma) + \mathbf{s}(\mathbf{y}) \log(\sigma) \tag{4.4}$$

e

$$\ell_2(\mu \mid \mathbf{y}) = (n - \mathbf{s}(\mathbf{y})) \log\left(\frac{\mu^2}{(1-\mu)} \right) + \mathbf{r}(\mathbf{y}) - \frac{\mu}{(1-\mu)} \mathbf{t}(\mathbf{y}), \tag{4.5}$$

em que $\mathbf{s}(\mathbf{y}) = \sum_{i=1}^{n} \delta_c(y_i)$, $\mathbf{r}(\mathbf{y}) = -3 \sum_{\substack{i=1 \\ y_i \in (0,1)}} \log(y_i)$ e $\mathbf{t}(\mathbf{y}) = \sum_{\substack{i=1 \\ y_i \in (0,1)}} \left(\frac{(1-y_i)}{y_i} \right)$.

Do ponto de vista frequentista, o estimador de máxima verossimilhança (EMV) $\widehat{\Psi} = (\widehat{\mu}, \widehat{\sigma})$ de Ψ pode ser obtido maximizando $\ell(\Psi \mid \mathbf{y})$ com respeito a σ e a μ. As derivadas parciais de primeira ordem de $\ell(\Psi \mid \mathbf{y})$ são dadas por

$$\frac{\partial}{\partial \sigma} \ell(\Psi \mid \mathbf{y}) = \frac{\mathbf{s}(\mathbf{y})}{\sigma} - \frac{n - \mathbf{s}(\mathbf{y})}{1-\sigma}$$

e

$$\frac{\partial}{\partial \mu} \ell(\Psi \mid \mathbf{y}) = \frac{(n - \mathbf{s}(\mathbf{y}))\mu^2 + (-3n + 3\mathbf{s}(\mathbf{y}) - \mathbf{t}(\mathbf{y}))\mu + 2(n - \mathbf{s}(\mathbf{y}))}{\mu(1-\mu)^2},$$

o que implica que o EMV $\widehat{\Psi}$ seja dado pelos componentes

$$\widehat{\sigma} = \frac{\mathbf{s}(\mathbf{y})}{n} \tag{4.6}$$

e

$$\widehat{\mu} = \frac{3(n - \mathbf{s}(\mathbf{y})) + \mathbf{t}(\mathbf{y}) - \sqrt{\mathbf{u}(\mathbf{y})}}{n - \mathbf{s}(\mathbf{y})}, \tag{4.7}$$

em que $\mathbf{u}(\mathbf{y}) = n^2 - 2\,n\,\mathbf{s}(\mathbf{y}) + 6\,n\,\mathbf{t}(\mathbf{y}) + \mathbf{s}^2(\mathbf{y}) - 6\,\mathbf{s}(\mathbf{y})\,\mathbf{t}(\mathbf{y}) + \mathbf{t}^2(\mathbf{y})$. O EMV do vetor de parâmetros Ψ das distribuições NULIZ e NULIU são obtidos substituindo c por zero e por um respectivamente nas Equações 4.6 e 4.7. É valido destacar que o estimador $\widehat{\mu}$ da

distribuição NUL inflacionada em c pode ser obtido diretamente pela expressão do EMV de μ da distribuição NUL para uma subamostra de Y com os extremos (zero e um) exclusos.

As derivadas parciais de segunda ordem de $\ell(\Psi \mid \mathbf{y})$ com respeito a σ e a μ são dadas respectivamente por

$$\frac{\partial^2}{\partial\sigma^2}\ell(\Psi \mid \mathbf{y}) = -\frac{\mathbf{s}(\mathbf{y})}{\sigma^2} - \frac{n-\mathbf{s}(\mathbf{y})}{(1-\sigma)^2}$$

e

$$\frac{\partial^2}{\partial\mu^2}\ell(\Psi \mid \mathbf{y}) = \frac{(-n+\mathbf{s}(\mathbf{y}))\mu^3 + (5[n-\mathbf{s}(\mathbf{y})] + 2\mathbf{t}(\mathbf{y}))\mu^2 + 6(n-\mathbf{s}(\mathbf{y}))\mu + 2(n-\mathbf{s}(\mathbf{y}))}{\mu^2(-1+\mu)^3},$$

o que condiciona a matriz de informação esperada de Fisher ser dada por

$$\mathbf{K}(\Psi) = \begin{bmatrix} k_{\mu\,\mu} & 0 \\ 0 & k_{\sigma\,\sigma} \end{bmatrix}$$

em que $k_{\mu\,\mu} = n\,(2-\mu^2)\,(1-\sigma)\,\mu^{-2}\,(1-\mu)^{-2}$ e $k_{\sigma\,\sigma} = n\,\sigma^{-1}\,(1-\sigma)^{-1}$. Deste modo, os intervalos de confiança e testes de hipóteses de $\Psi = (\mu, \sigma)$, podem ser construídos a partir do pressuposto de normalidade assintótica dos estimadores de máxima verossimilhança (CASELLA; BERGER, 2002).

5 Modelo de Regressão New unit-Lindley inflacionado em Zero ou em Um

Nesta seção estabelecemos uma estrutura de regressão para as distribuições NULIZ e NULIU a partir de um termo mais geral dada pela Equação 3.1. Embora estes modelos possam ser incorporados aos modelos lineares generalizados, uma vez que pertencem a família exponencial, optamos por definir os Modelos de Regressão New unit-Lindley inflacionado em Zero (MRNULIZ) ou em Um (MRNULIU) como membros de uma família mais geral conhecida como Modelos Generalizados Aditivos para Locação, Escala e Forma , do inglês, Generalised Additive Models for Location Scale and Shape (GAMLSS) (RIGBY; STASINOPOULOS, 2005). Visto isso, e tendo conhecimento da natureza continua-discreta destas distribuições, optamos pela parametrização $\Psi = (\mu, \sigma)$ que permite modelar separadamente os componentes contínuos e discreto destes modelos. Nessas condições,

Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes tais que, para todo $i = 1, \ldots, n$, Y_i segue uma distribuição NUL inflacionada em c com média condicional μ_i e massa de probabilidade σ_i. Nessas condições, os modelo de regressão NUL inflacionado em c é definido pelos componentes sistemáticos

$$h_1(\mu_i) = \sum_{j=1}^{n_1} u_{ij}^{\intercal}\beta_j = \eta_i \text{ e } h_2(\sigma_i) = \sum_{j=1}^{n_2} v_{ij}^{\intercal}\gamma_j = \zeta_i,$$

em que $u_i = (u_{i1}, u_{i2}, \ldots, u_{in_1})$ e $v - i = (v_{i1}, v_{i2}, \ldots, v_{in_2})$ são o conjunto de covariáveis fixas $(n_1 + n_2 < n)$, os vetores $\beta = (\beta_1, ..., \beta_{n_1})$ e $\gamma = (\gamma_1, ..., \gamma_{n_2})$ são coeficientes regressores desconhecidos e, $h_1 : (0,1) \mapsto (-\infty, \infty)$ e $h_2 : (0,1) \mapsto (-\infty, \infty)$ são funções de ligação continuas, estritamente monótonas e duas vezes diferenciáveis em μ e σ. Dentre as potenciais especificações de $h_1(\cdot)$ e $h_2(\cdot)$, temos as funções logito, probito, log-log e a log-log complementar como as mais usuais. Sob a perspectiva clássica, os estimadores dos coeficientes β e γ são soluções, respectivamente, das equações

$$0 = \frac{\partial}{\partial \sigma_i} \dot{\ell}_1(\gamma \mid \mathbf{y}) \frac{\partial}{\partial \gamma_j} \sigma_i$$

e

$$0 = \frac{\partial}{\partial \mu_i} \dot{\ell}_2(\beta \mid \mathbf{y}) \frac{\partial}{\partial \beta_j} \mu_i$$

em que

$$\dot{\ell}_1(\gamma \mid \mathbf{y}) = \sum_{i=1}^{n} (1 - \delta_c(y_i)) \log(1 - \sigma_i) + \delta_c(y_i) \log(\sigma_i)$$

e

$$\dot{\ell}_2(\beta \mid \mathbf{y}) = \sum_{\substack{i=1 \\ y_i \in (0,1)}} 2\log(\mu_i) - \log(1 - \mu_i) - \frac{(1 - y_i)}{y_i} \mu_i (1 - \mu_i)^{-1}.$$

6 Agregando Zeros e Uns à distribuição New unit-Lindley

Para agregar zeros e uns à distribuição NUL, recorremos ao modelo de mistura com a distribuição de Bernoulli atribuindo assim probabilidades positivas aos extremos do intervalo. Deste modo, as funções de distribuição acumulada e densidade de probabilidade desta mistura são dadas respectivamente por

$$\text{FNULIZU}(y \mid \mu, \sigma, \rho) = \sigma\, \text{Ber}(y \mid \rho) + (1 - \sigma)\text{F}(y \mid \mu), \tag{6.1}$$

e

$$\text{fnulizu}(y \mid \mu, \sigma, \rho) = \sigma\, \text{ber}(y \mid \rho) + (1 - \sigma)\text{f}(y \mid \mu) \tag{6.2}$$

em que $\text{Ber}(\cdot \mid \rho)$ e $\text{ber}(\cdot \mid \rho)$ denotam respectivamente a f.d.a. e f.d.p. de uma distribuição de Bernoulli com probabilidade de sucesso ρ enquanto que, $\text{F}(\cdot \mid \mu)$ e $\text{f}(\cdot \mid \mu)$, são a f.d.a. e f.d.p. NUL, dadas respectivamente, pelas Equações 2.3 e 2.4.

Definição 2 *Uma variável aleatória Y é dito ser uma distribuição New unit-Lindley Inflacionada em Zero e Um, denotada por $Y \sim$ NULIZU(μ, σ, ρ), se suas f.d.a. e f.d.p. foram expressas pelas Equações 6.1 e 6.2.*

Alternativamente a Equação 6.2, podemos expressar a f.d.p. NULIZU como uma função definida por ramos dada por

$$\text{fnulizu}(y \mid \mu, \sigma, \rho) = \begin{cases} \sigma(1-\rho) & \text{se } y = 0 \\ (1-\sigma)\text{f}(y \mid \mu) & \text{se } 0 < y < 1 \\ \sigma\rho & \text{se } y = 1 \end{cases}, \tag{6.3}$$

em que os valores $\sigma(1-\rho)$ e $\sigma\rho$ correspondem respectivamente as probabilidades de se observarem zeros e uns. Por outro lado, se $Y \sim$ NULIZU(μ, σ, ρ), então $\mathbb{E}[Y^r] = \mu_r^{'} = \mathbb{E}[(1-\sigma)\, X^r + \sigma\, W^r] = (1-\sigma)\, \mathbb{E}[X^r] + \sigma\, \mathbb{E}[W^r] = (1-\sigma)\, \mu_r + \sigma\rho$, em que $r = 1, 2, 3, \ldots$ corresponde a ordem do momento avaliado, X uma variável aleatória NUL e W uma variável aleatória Bernoulli. Em particular, para $r = 1, 2$ temos que a média e variância do modelo NULIZ são dadas respectivamente por

$$\begin{aligned} \mu_1^{'} &= (1-\sigma)\, \mu + \sigma\rho \\ \mu_2^{'} &= (1-\sigma) \left[\mu^2\, \mathbf{d}(\mu)(1-\mu)^{-1}\right] + \sigma\rho \end{aligned}$$

em que $\mathbf{d}(\mu) = \text{Ei}[1,\, \mu(1-\mu)^{-1}]\, \exp\left(\mu(1-\mu)^{-1}\right)$ e $\text{Ei}[a, z] = \int_1^\infty z^{-a} e^{-yz} dz$ é a função integral exponencial (ABRAMOWITZ; STEGUN, 1974).

Da mesma forma que as distribuições NULIZ e NULIU, a distribuição NULIZU é um modelo de mistura com natureza contínua e discreta e conta com pontos de descontinuidade nas massas de probabilidade em zero e um. Sua densidade, com suporte ampliado nos pontos zero e um, pode assumir diversas formas sendo multimodal quando $\mu < 3/4$, $\sigma = \mathbf{b}(\mu)/(\mathbf{b}(\mu) + \mu/2)$ e $\rho = 1/2$ em que $\mathbf{b}(\mu) = 27 \exp\left[(4\mu - 3)(1-\mu)^{-1}\right] (\mu - 1)^2$.

Proposição 2 *A distribuição NULIZU pertence a família exponencial multiparamétrica.*

Prova: De fato, sendo $\eta = (\eta_1, \eta_2, \eta_3)$, com $\eta_1 = \log(\sigma(1-\rho)) - \log(1-\sigma) - \log\left(\dfrac{\mu^2}{1-\mu}\right)$, $\eta_2 = \log(\sigma\rho) - \log(1-\sigma) - \log\left(\dfrac{\mu^2}{1-\mu}\right)$ e $\eta_3 = -\dfrac{\mu}{1-\mu}$; e $T(y) = (t_1(y), t_2(y), t_3(y))$, com $t_1(y) = \delta_0(y)$, $t_2(y) = \delta_1(y)$ e $t_3(y) = \dfrac{1-y}{y}$ se $y \in (0, 1)$ e igual a 0 caso contrario, obtém-se a Equação 6.3 escrita na forma

$$\text{fnulizu}(y \mid \mu, \sigma, \rho) = \exp\{\eta^\top T(y) - \zeta(\eta)\}\, h(y)\,,$$

em que $\zeta(\eta) = \eta_2 - \log(\sigma\rho)$ e $h(y) = y^{-3}$ se $y \in (0,1)$ e igual a 1 caso contrario. ■

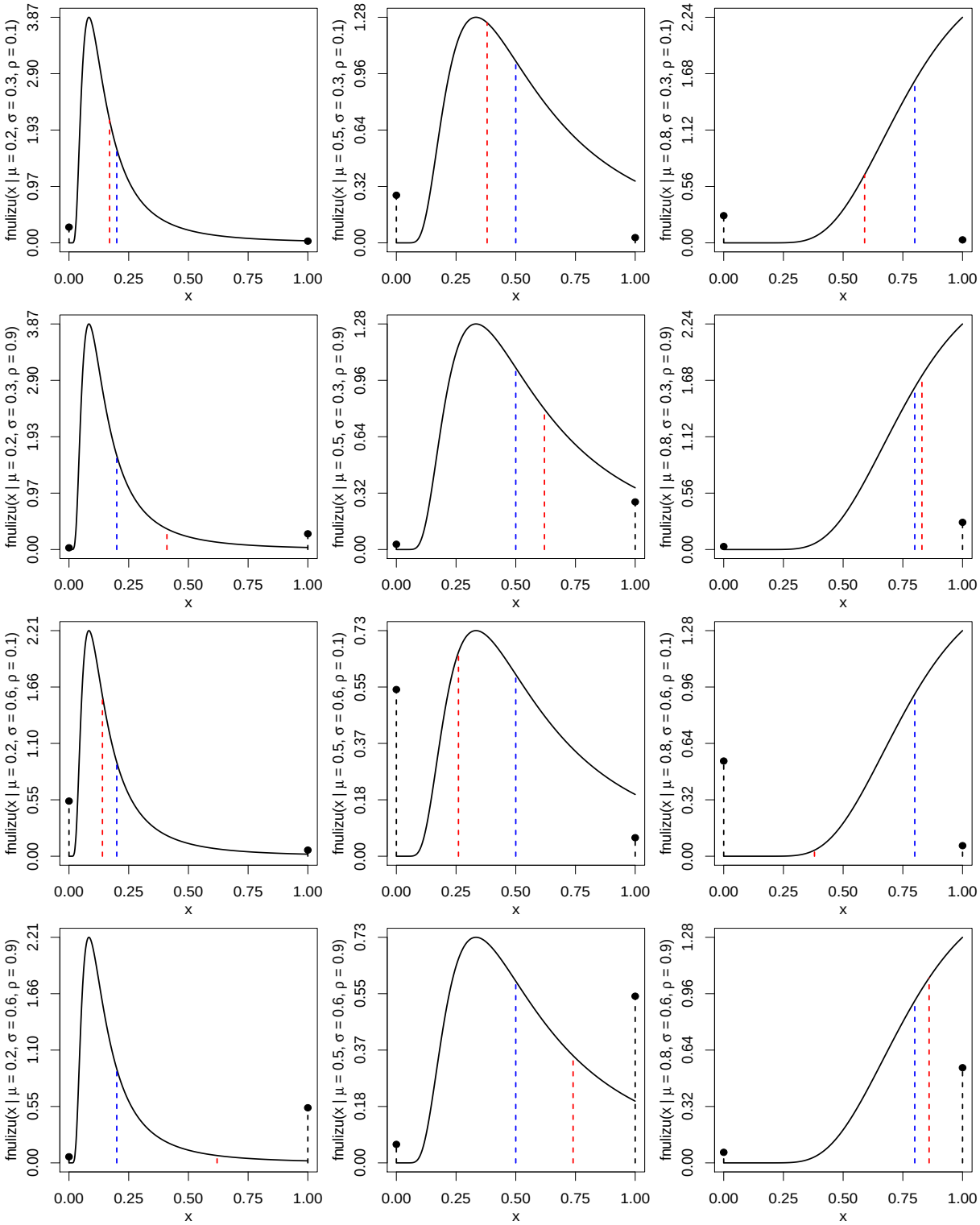

Figura 6.1 – Comportamento da densidade NULIZU assumindo diferentes valores para o vetor paramétrico (μ, σ, ρ).(linha pontinhada em preto: proporção de zeros ou uns; linha pontilhada em azul: média da distribuição New unit-Lindley; linha pontilhada em vermelho: média conjunta do modelo de mistura)

A distribuição NULIZU conta também com expressão fechada para a função quantil. Posto isso, para $0 < p < 1$, os respectivos quantis são dados por

$$QNULIZU(p \mid \mu, \sigma, \rho) = \begin{cases} 0 & \text{se } \sigma(1-\rho) \geq p \\ 1 & \text{se } (1-\sigma\rho) \leq p \\ \dfrac{\mu}{\left(W_{-1}\left[-\dfrac{(p-\sigma(1-\rho))\,\mathbf{e}(\mu)}{(1-\sigma)}\right]+1\right)(-1+\mu)} & \text{caso contrario} \end{cases}$$

em que $\mathbf{e}(\mu) = (1-\mu)^{-1}\exp(-(1-\mu)^{-1})$ e W_{-1} denota o ramo negativo da função `W` de Lambert (CORLESS et al., 1996; JODRÁ, 2010).

7 Aspectos Inferências da distribuição New unit-Lindley inflacionada em Zero e Um

Sejam $Y = Y_1, Y_2. \ldots, Y_n$ uma amostra aleatória retirada de uma população com f.d.p. dada pela Equação 6.3. Sem perca de generalidade, densidade da variável Y pode ser expressa na forma

$$\text{fnulizu}(y \mid \mu, \sigma, \rho) = \{(\sigma(1-\rho))^{\delta_0(y)}\,(\sigma\rho)^{\delta_1(y)}(1-\sigma)^{1-\delta_{\{0,1\}}(y)}\}\{\text{f}(y \mid \mu)\}^{1-\delta_{\{0,1\}}(y)},$$

em que fnulizu$(y \mid \mu, \sigma, \rho)$, como função dos parâmetros, pode ser fatorado em dois termos independentes sendo o primeiro deles em função de σ e ρ e o segundo termo em função de μ. Nessas condições, para $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ observado, a função de verossimilhança de $\Theta = (\mu, \sigma, \rho)$ é dada por

$$L(\Theta \mid \mathbf{y}) = L_1(\sigma, \rho \mid \mathbf{y}) \times L_2(\mu \mid \mathbf{y}), \tag{7.1}$$

em que

$$L_1(\sigma, \rho \mid \mathbf{y}) = \prod_{i=1}^{n}(\sigma(1-\rho))^{\delta_0(y_i)}\,(\sigma\rho)^{\delta_1(y_i)}(1-\sigma)^{1-\delta_{\{0,1\}}(y_i)}$$

e

$$L_2(\mu \mid \mathbf{y}) = \prod_{i=1}^{n}\left\{\frac{\mu^2}{(1-\mu)\,y_i^3}\exp\left(-\frac{\mu\,(1-y_i)}{(1-\mu)\,y_i}\right)\right\}^{1-\delta_{\{0,1\}}(y_i)}.$$

O logaritmo da Equação 7.1 é dado por

$$\ell(\Theta \mid \mathbf{y}) = \ell_1(\sigma, \rho \mid \mathbf{y}) + \ell_2(\mu \mid \mathbf{y}), \tag{7.2}$$

em que

$$\ell_1(\sigma,\rho \mid \mathbf{y}) = \mathbf{s_0}(\mathbf{y})\log(\sigma(1-\rho)) + \mathbf{s_1}(\mathbf{y})\log(\sigma\rho) + (n - \mathbf{s_0}(\mathbf{y}) - \mathbf{s_1}(\mathbf{y}))\log(1-\sigma) \quad (7.3)$$

$$\ell_2(\mu \mid \mathbf{y}) = (n - \mathbf{s_0}(\mathbf{y}) - \mathbf{s_1}(\mathbf{y}))\log\left(\frac{\mu^2}{(1-\mu)}\right) + \mathbf{r}(\mathbf{y}) - \frac{\mu}{(1-\mu)}\mathbf{t}(\mathbf{y}) \quad (7.4)$$

com $\mathbf{s_0}(\mathbf{y}) = \sum_{i=1}^{n} \delta_0(y_i)$, $\mathbf{s_1}(\mathbf{y}) = \sum_{i=1}^{n} \delta_1(y_i)$, $\mathbf{r}(\mathbf{y}) = -3 \sum_{\substack{i=1 \\ y_i \in (0,1)}} \log(y_i)$ e $\mathbf{t}(\mathbf{y}) = \sum_{\substack{i=1 \\ y_i \in (0,1)}} \left(\frac{1-y_i}{y_i}\right)$.

Posto isto, o EMV $\widehat{\Theta} = (\widehat{\mu}, \widehat{\sigma}, \widehat{\rho})$ de Θ é dado por

$$\begin{aligned}
\widehat{\mu} &= \frac{1}{2\,\dot{\mathbf{s}}(\mathbf{y})}\left(3\,\dot{\mathbf{s}}(\mathbf{y}) + \mathbf{t}(\mathbf{y}) + \sqrt{\mathbf{v}(\mathbf{y})}\right), \\
\widehat{\sigma} &= \frac{\mathbf{s_0}(\mathbf{y}) + \mathbf{s_1}(\mathbf{y})}{n}, \\
\widehat{\rho} &= \frac{\mathbf{s_1}(\mathbf{y})}{\mathbf{s_0}(\mathbf{y}) + \mathbf{s_1}(\mathbf{y})},
\end{aligned}$$

em que $\mathbf{v}(\mathbf{y}) = n^2 + 2n\,[3\mathbf{t}(\mathbf{y}) - \mathbf{s_0}(\mathbf{y}) - \mathbf{s_1}(\mathbf{y})] + \mathbf{t}(\mathbf{y})^2 - 6\mathbf{t}(\mathbf{y})\,[\mathbf{s_0}(\mathbf{y}) + \mathbf{s_1}(\mathbf{y})] + [\mathbf{s_0}(\mathbf{y}) + \mathbf{s_1}(\mathbf{y})]^2$ e $\dot{\mathbf{s}}(\mathbf{y}) = n - \mathbf{s_0}(\mathbf{y}) - \mathbf{s_1}(\mathbf{y})$.

As derivadas de segunda ordem da Equação 7.2 em relação ao vetor de parâmetros Θ são dadas por

$$\begin{aligned}
\frac{\partial^2}{\partial\mu^2}\,\ell(\Theta \mid \mathbf{y}) &= \frac{\mu^3\,\dot{\mathbf{s}}(\mathbf{y}) - \mu^2\,[5\,\dot{\mathbf{s}}(\mathbf{y}) + 2\,\mathbf{t}(\mathbf{y})] + 6\,\mu\,\dot{\mathbf{s}}(\mathbf{y}) - 2\,\dot{\mathbf{s}}(\mathbf{y})}{\mu^2\,(1-\mu)^3}, \\
\frac{\partial^2}{\partial\sigma^2}\,\ell(\Theta \mid \mathbf{y}) &= \frac{-n\,\sigma^2 + 2\,\sigma\,[\mathbf{s_0}(\mathbf{y}) + \mathbf{s_1}(\mathbf{y})] - \mathbf{s_0}(\mathbf{y}) - \mathbf{s_1}(\mathbf{y})}{\sigma^2\,(1-\sigma)^2}, \\
\frac{\partial^2}{\partial\rho^2}\,\ell(\Theta \mid \mathbf{y}) &= -\frac{\mathbf{s_0}(\mathbf{y})}{(1-\rho)^2} - \frac{\mathbf{s_1}(\mathbf{y})}{\rho^2}, \\
\frac{\partial^2}{\partial\mu\,\partial\sigma}\,\ell(\Theta \mid \mathbf{y}) &= 0, \\
\frac{\partial^2}{\partial\mu\,\partial\rho}\,\ell(\Theta \mid \mathbf{y}) &= 0, \\
\frac{\partial^2}{\partial\sigma\,\partial\rho}\,\ell(\Theta \mid \mathbf{y}) &= 0.
\end{aligned}$$

Desse modo, a matriz de informação esperada de Fisher é dada por

$$\mathbf{K}(\Theta) = \begin{bmatrix} k_{\mu\,\mu} & 0 & 0 \\ 0 & k_{\sigma\,\sigma} & 0 \\ 0 & 0 & k_{\rho\,\rho} \end{bmatrix}$$

em que $k_{\mu\,\mu} = n\,(2-\mu^2)\,(1-\sigma)\,\mu^{-2}\,(1-\mu)^{-2}$, $k_{\sigma\,\sigma} = n\,\sigma^{-1}\,(1-\sigma)^{-1}$ e $k_{\rho\,\rho} = n\,\sigma\,\rho^{-1}\,(1-\rho)^{-1}$. Os intervalos de confiança e testes de hipóteses de $\Theta = (\mu, \sigma, \rho)$ são construídos a partir do pressuposto de normalidade assintótica dos estimadores de máxima verossimilhança (CASELLA; BERGER, 2002).

8 Modelo de Regressão New unit-Lindley inflacionada em Zero e Um

Nesta seção estabelecemos uma estrutura de regressão para distribuição NULIZU. Tal como as distribuições NULIZ e NULIU, a distribuição NULIZU pode ser incorporado aos modelos lineares generalizados porém nesta seção definimos o modelo de regressão New unit-Lindley inflacionado em Zero e Um (MRNULIZU) como membro da família GAMLSS.

Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes tais que, para todo $i = 1, \ldots, n$, Y_i segue uma distribuição NULIZU com parâmetros μ_i, σ_i e ρ_i. Nessas condições, o MRNULIZU é definido pelos componentes sistemáticos

$$h_1(\mu_i) = \sum_{j=1}^{n_1} u_{ij}\beta_j = \eta_i,\ h_2(\sigma_i) = \sum_{j=1}^{n_2} v_{ij}\gamma_j = \zeta_i \text{ e } h_3(\rho_i) = \sum_{j=1}^{n_3} w_{ij}\lambda_j = \pi_i,$$

em que $u_i = (u_{i1}, u_{i2}, \ldots, u_{in_1})$, $v_i = (v_{i1}, v_{i2}, \ldots, v_{in_2})$ e $w_i = (w_{i1}, w_{i2}, \ldots, w_{in_3})$ são o conjunto de covariáveis fixas $(n_1+n_2+n_3 < n)$, os vetores $\beta = (\beta_1, ..., \beta_{n_1})$, $\gamma = (\gamma_1, ..., \gamma_{n_2})$ e $\lambda = (\lambda_1, ..., \lambda_{n_3})$ são coeficientes regressores desconhecidos e, $h_1 : (0,1) \mapsto (-\infty, \infty)$, $h_2 : (0,1) \mapsto (-\infty, \infty)$ e $h_3 : (0,1) \mapsto (-\infty, \infty)$ são funções de ligação continuas, estritamente monótonas e duas vezes diferenciáveis em μ, σ e ρ.

Novamente sobre uma abordagem clássica, temos que o estimador de β é solução da equação

$$0 = \frac{\partial}{\partial \mu_i}\ddot{\ell}_2(\beta \mid \mathbf{y})\frac{\partial}{\partial \beta_j}\mu_i$$

enquanto que, os estimadores de γ e λ são soluções do sistema de equações

$$0 = \frac{\partial}{\partial \sigma_i}\ddot{\ell}_1(\gamma, \lambda \mid \mathbf{y})\frac{\partial}{\partial \gamma_j}\sigma_i$$
$$0 = \frac{\partial}{\partial \rho_i}\ddot{\ell}_1(\gamma, \lambda \mid \mathbf{y})\frac{\partial}{\partial \lambda_j}\rho_i$$

em que

$$\ddot{\ell}_1(\gamma, \lambda \mid \mathbf{y}) = \sum_{i=1}^{n} \delta_0(y_i)\log(\sigma_i(1-\rho_i)) + \delta_1(y_i)\log(\sigma_i\rho_i) + (1 - \delta_{\{0,1\}}(y_i))\log(1-\sigma_i)$$

e

$$\ddot{\ell}_2(\beta \mid \mathbf{y}) = \sum_{\substack{i=1 \\ y_i \in (0,1)}} 2\log(\mu_i) - \log(1-\mu_i) - \frac{(1-y_i)}{y_i}\mu_i(1-\mu_i)^{-1}.$$

9 Estudo de Simulação

Nesta seção conduzimos um estudo de simulação Monte Carlo para avaliar a performance dos coeficientes regressores dos modelos de regressão NUL com inflação em zero, em

um e em zero e um. O experimento foi conduzido com base em 10000 repetições para os tamanhos de amostra $n = 50,\ 100,\ 150, \ldots,\ 300$ das quais foram avaliados o viés e erro quadrático médio (EQM) dados respectivamente por

$$\text{Viés}\,(\widehat{\xi}) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\xi}_i - \xi) \quad \text{e} \quad \text{EQM}\,(\widehat{\xi}) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\xi}_i - \xi)^2$$

em que ξ denota o coeficiente avaliado em $N = 10000$ simulações.

Consideramos como cenários, os modelos de regressão múltipla (com duas covariáveis) sobre todos os parâmetros da distribuição e sobre o parâmetro de média condicional mantendo as massas de probabilidades fixas. Escolhemos a especificação logito para as funções de ligação de modo que ficamos com os componentes sistemáticos

$$\begin{aligned}
&\text{logito}(\mu_i) = \beta_0 + \beta_1 u_{i1} + \beta_2 u_{i2}\\
&\text{logito}(\sigma_i) = \gamma_0 + \gamma_1 u_{i1} + \gamma_2 u_{i2}\\
&\text{e}\\
&\text{logito}(\rho_i) = \lambda_0 + \lambda_1 u_{i1} + \lambda_2 u_{i2}
\end{aligned}$$

em que u_{i1} e u_{i2} são elementos do vetor de variáveis gerados a partir de uma distribuição uniforme, no -1 à 1, e os coeficientes da regressão dados por $\beta = (\beta_0, \beta_1, \beta_2) = \delta = (\delta_0, \delta_1, \delta_2) = \lambda = (\lambda_0, \lambda_1, \lambda_2) = (1, -2, 3)$.

Os resultados deste estudo são apresentados da seguinte maneira. As Figuras 9.1, 9.2, 9.3 e 9.4 correspondem a performance da simulação no MRNULIZ, As Figuras 9.5, 9.6, 9.7 e 9.8 a performance do MRNULIU e as Figuras 9.9, 9.10, 9.11 e 9.12 correspondem ao MRNULIZU. Neste estudo, observamos que

- Em todos os cenários, os viéses e EQM's apresentam tendem para zero a medida que o tamanho da amostra n aumenta.
- Nos modelos de regressão sobre a média condicional μ mantendo fixos os outros parâmetros, Figuras 9.3, 9.4, 9.7, 9.8, 9.11 e 9.12, o viés e EQM de $\widehat{\beta}$ aumenta a medida que σ também aumenta. Isso possivelmente ocorre porque valores mais altos de σ implica em subamostras menores de elementos de uma variável NUL e consequentemente uma quantidade menor de informação para determinar os estimadores.
- Nos modelos de regressão sobre todos os parâmetros, o intercepto geralmente apresenta o menor valor de EQM. Veja as Figuras 9.2, 9.6 e 9.10.
- O EQM de $\widehat{\beta}$ é geralmente menor do que o de $\widehat{\gamma}$ nos tamanhos de amostras aqui considerados. Veja as Figuras 9.2, 9.6 e 9.10.

- No MRNULIZU, com σ e ρ fixos, o parâmetro ρ não apresenta influência significativa sobre o viés e EQM de $\widehat{\beta}$. Veja as Figuras 9.11 e 9.12.

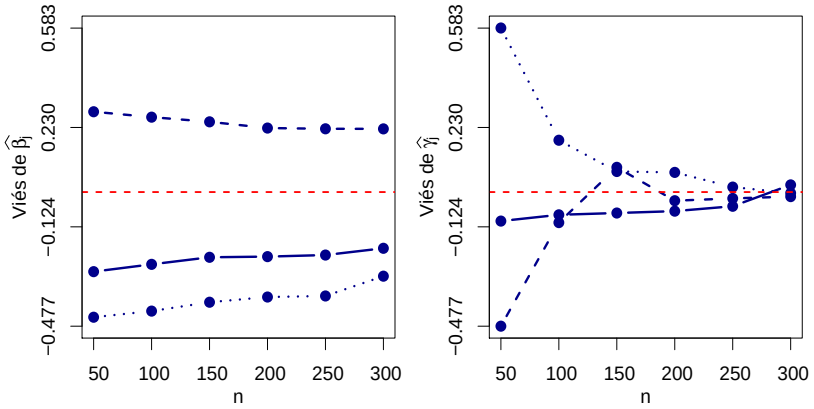

Figura 9.1 – Viés das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZ (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2).

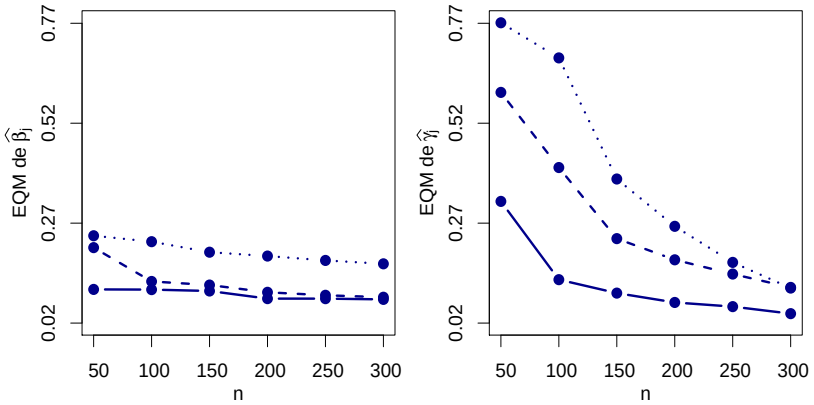

Figura 9.2 – EQM das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZ (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2).

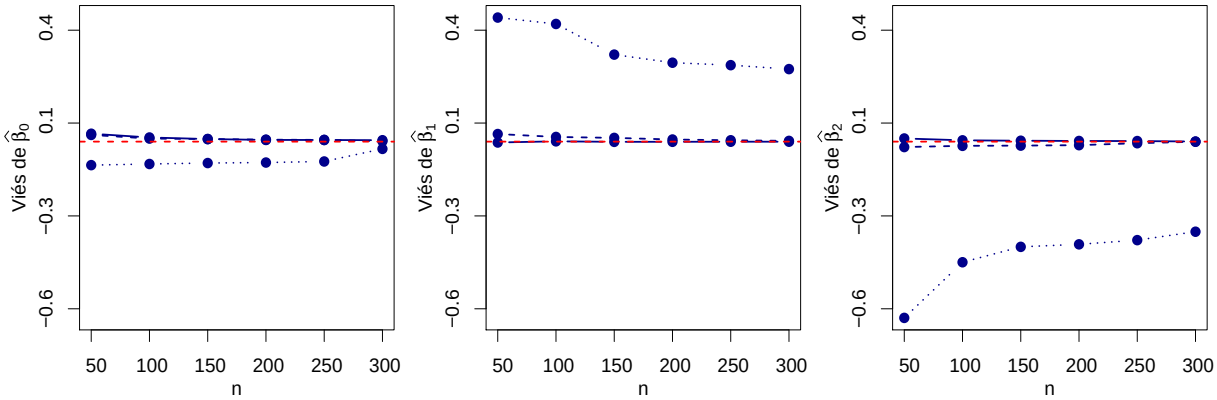

Figura 9.3 – Viés das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZ mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$).

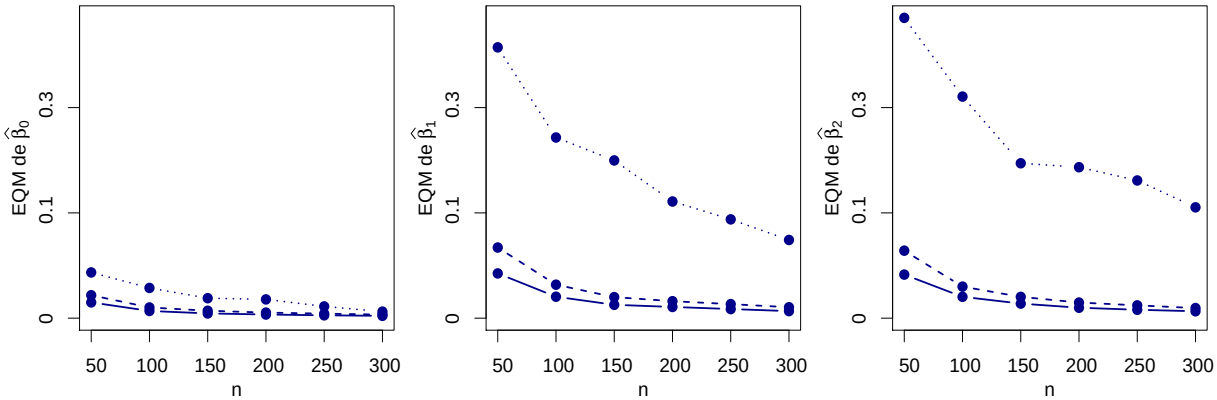

Figura 9.4 – EQM das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZ mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$).

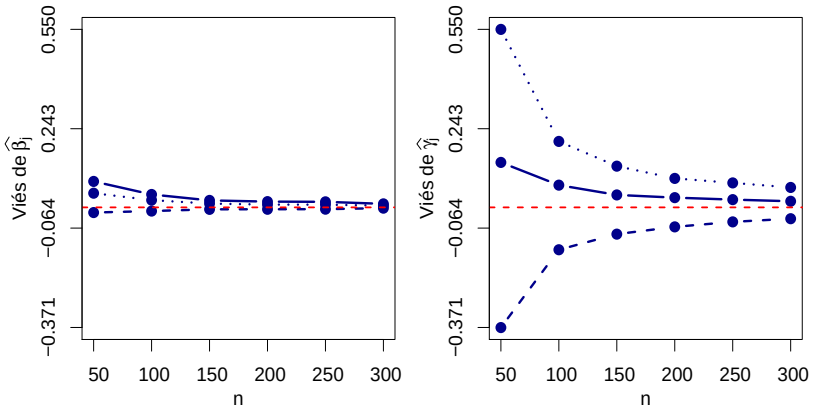

Figura 9.5 – Viés das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2).

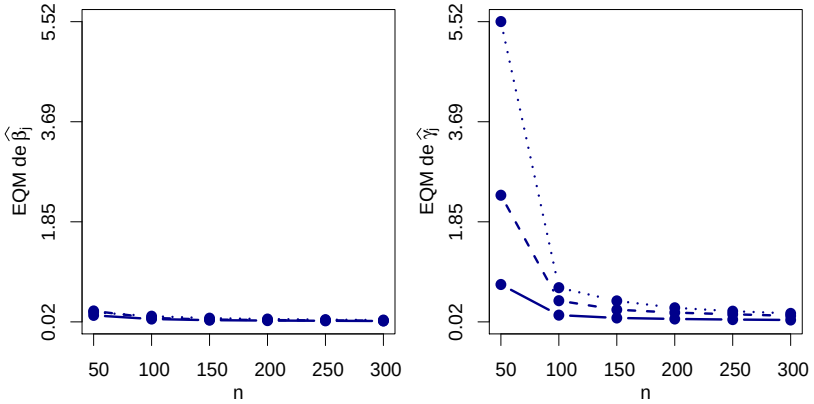

Figura 9.6 – EQM das estimativas $\widehat{\beta}$ e $\widehat{\gamma}$ dos coeficientes regressores do vetor de parâmetros no MRNULIU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2).

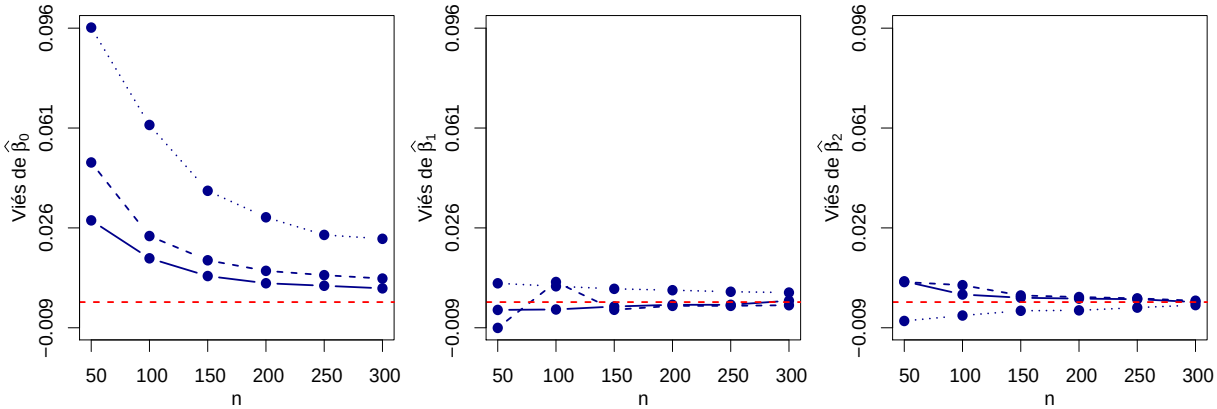

Figura 9.7 – Viés das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIU mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$).

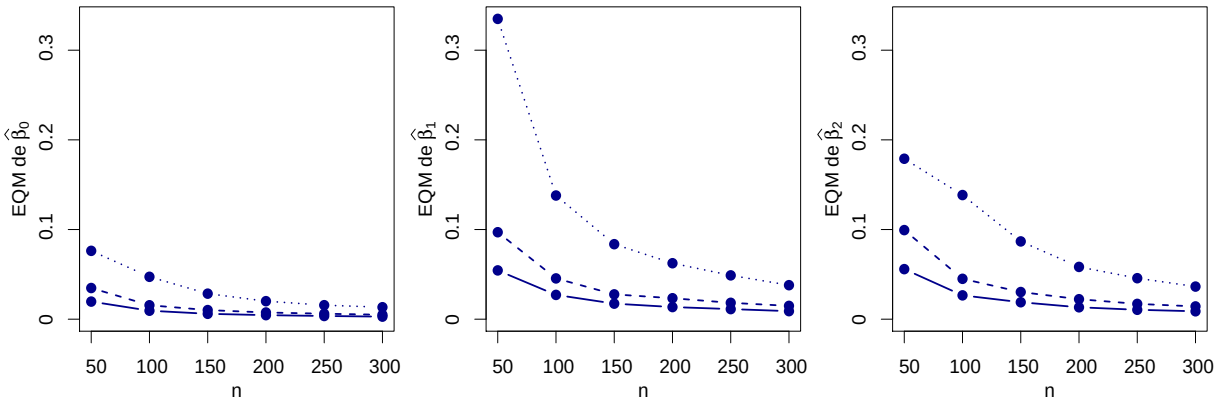

Figura 9.8 – EQM das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIU mantendo σ fixo (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.5$ e linha pontilhada: $\sigma = 0.8$).

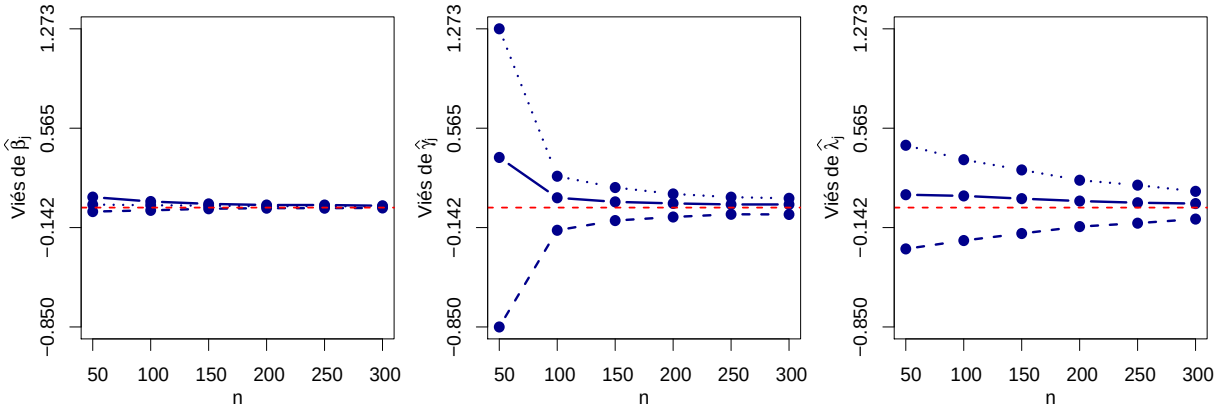

Figura 9.9 – Viés das estimativas $\widehat{\beta}$, $\widehat{\gamma}$ e $\widehat{\lambda}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2).

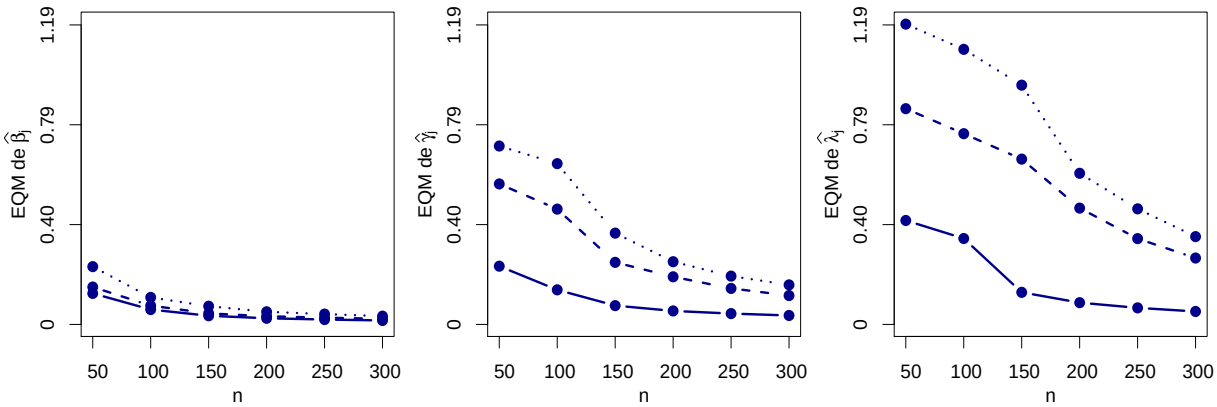

Figura 9.10 – EQM das estimativas $\widehat{\beta}$, $\widehat{\gamma}$ e $\widehat{\lambda}$ dos coeficientes regressores do vetor de parâmetros no MRNULIZU (linha contínua: j=0, linha tracejada: j=1 e linha pontilhada: j=2).

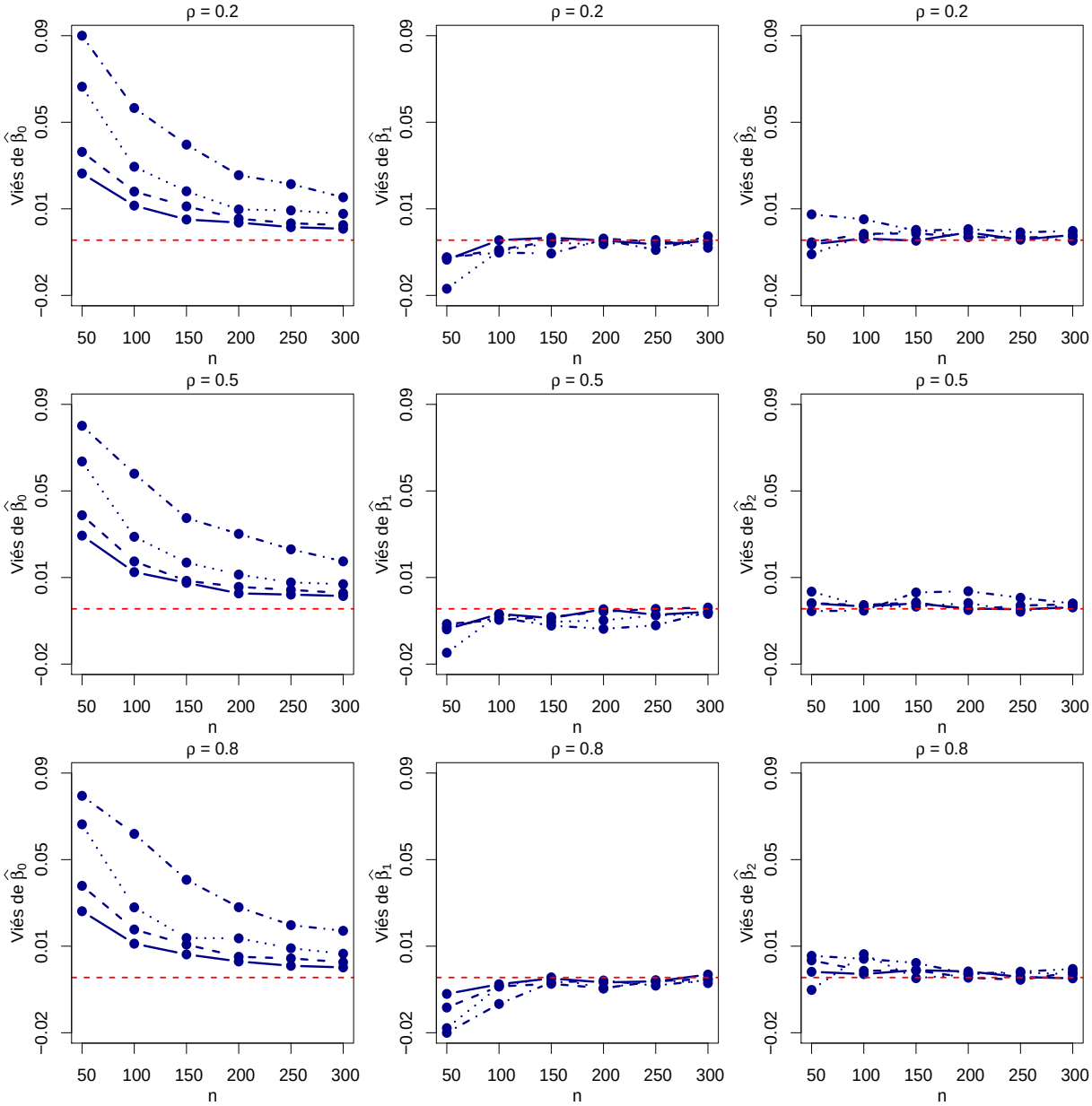

Figura 9.11 – Viés das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZU mantendo σ e ρ fixos (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.4$, linha pontilhada: $\sigma = 0.6$ e linha tracejada-pontilhada: $\sigma = 0.8$).

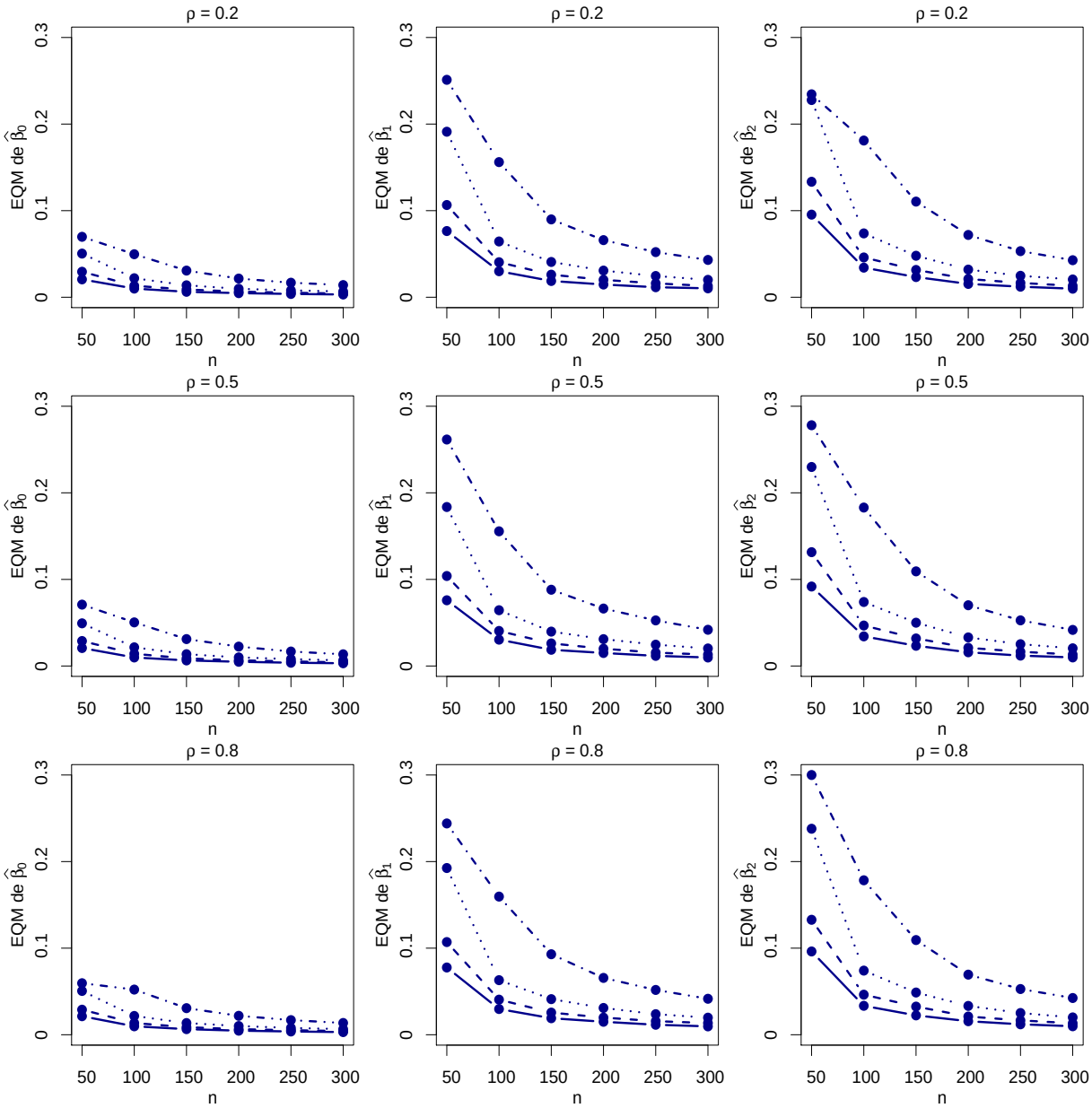

Figura 9.12 – EQM das estimativas dos coeficientes regressores de μ, $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2})$, no MRNULIZU mantendo σ e ρ fixos (linha contínua: $\sigma = 0.2$, linha tracejada: $\sigma = 0.4$, linha pontilhada: $\sigma = 0.6$ e linha tracejada-pontilhada: $\sigma = 0.8$).

10 Aplicação a Dados Reais

Nesta seção, adotamos o MRNULIZ para descrever o comportamento da taxa anual de suicídio no ano de 2016 para países dos continentes africano, asiático e europeu levando em conta informações a respeito do sexo, da faixa-etária e de índices socioeconômicos do país como a taxa de desemprego, o produto interno bruto (PIB) e índice de desenvolvimento humano (IDH). Os dados utilizados nesta análise possuem domínio público e estão disponíveis

nos endereços eletrônicos <https://www.who.int/>, <http://hdr.undp.org/en> e <https://databank.worldbank.org/source/world-development-indicators>.

Na Tabela 1, disponibilizamos uma breve descrição das variáveis de natureza numérica sendo o PIB dado na escala logarítmica (logaritmo neperiano). Note que, comparada com as outras variáveis, a taxa de suicídio é dimensionalmente muito menor o que nos levou a trabalhar com o PIB na escala logarítmica. No total, contamos com 140 observações sobre a taxa de suicídio em grupos de indivíduos das faixa etárias de 15 à 24 anos (idade 0); de 25 à 34 anos (idade 1); de 35 à 54 anos(idade 2); de 55 à 74 anos (idade 3) desagregados por sexo (feminino: 0 e masculino: 1). Em sua maioria, povos asiáticos e europeus constituem cerca de 93% das observações dispostas neste banco de dados.

Tabela 1 – Resumo estatístico das variáveis de natureza numérica

	Taxa de suicídio	log(PIB)	Taxa de desemprego	IDH
Mínimo	0	23,08	$0,15$	0,73
1º Quartil	$2,5 \times 10^{-5}$	23,23	$6,01$	0,77
Mediana	$7,3 \times 10^{-5}$	24,57	$7,12$	0,84
Média	$1,2 \times 10^{-4}$	24,90	$8,71$	0,84
3º Quartil	$1,5 \times 10^{-4}$	26,69	$12,95$	0,87
Máximo	$9,7 \times 10^{-4}$	27,38	$17,62$	0,94

Para fins de comparação consideramos também os modelos de regressão unit-Lindley inflacionada em zero (MRULIZ) e Beta inflacionado em zero (MRBEIZ). Os ajustes e cálculos afins foram realizados por meio de funções dispostas nas bibliotecas `gamlss` (RIGBY; STASINOPOULOS, 2005) e `gamlss.inf` (ENEA et al., 2019) no software `R` (R Core Team, 2019) (Veja o Apêndice A para mais detalhes). Consideramos a especificação logito para as funções de ligação e, os modelos de regressão para os parâmetros μ_i, σ_i e ϕ_i são dados respectivamente pelos componentes

$$\text{logito}(\mu_i) = \beta_0 + \beta_1 \times \text{sexo } 1_i + \beta_2 \times \text{idade } 1_i + \beta_3 \times \text{idade } 2_i + \beta_4 \times \text{idade } 3_i + \beta_5 \times \text{idade } 4_i + \beta_6 \times \text{desemprego}_i + \beta_7 \times \text{IDH}_i$$

e

$$\text{logito}(\sigma_i) = \gamma_0 + \gamma_1 \times \text{sexo } 1_i + \gamma_2 \times \log(\text{PIB}_i) + \gamma_3 \times \text{desemprego}_i$$

Na Tabela 3 dispomos uma descrição detalhada com as estimativas, erros padrão e intervalos de confiança dos parâmetros ajustados. Observe que o conjunto de variáveis presentes nesta análise não foram significativas, ao nível de 5%, para explicar a média condicional do MRBEIZ. Por outro lado, as variáveis sexo, faixa etária, taxa de desemprego, e IDH foram

significativas, ao nível de 5%, para o MRNULIZ e MRULIZ. Por outro lado, no que diz respeito a massa de probabilidade em zero, as estimativas são as mesmas para os três modelos haja vista que eles possuem a mesma estrutura de inflacionamento. Tivemos como variáveis significativas ao nível de 5% neste caso, o sexo e a taxa de desemprego que foi significativa apenas na presença da variável PIB.

Na Tabela 2 apresentamos os critérios de informação Akaike (AIC) e Bayesiano de Schwarz (BIC) utilizados posteriormente para discriminação dos modelos. Observe que o MRNULIZ apresenta o menor valor (melhor) para os critérios considerados, contudo estes valores aparentemente não apresentam uma diferença considerável quando comparados com o AIC e BIC do MRULIZ. Deste modo, uma análise residual sobre cada modelo nos permitiu medir e determinar o modelo com melhor ajuste sobre os dados.

Tabela 2 – Estatísticas de discriminação de modelos

Modelo	AIC	BIC
MRNULIZ	-2119,50	-2084,20
MRULIZ	-2094,20	-2058,90
MRBEIZ	-1817,03	-1778,79

Na Figura 10.1 apresentamos a construção dos gráficos tipo *worm plot* que avaliam o desvio dos resíduos em relação aos quantis de uma distribuição normal padrão. No cenário ideal, os desvios devem seguir o mais próximo possível da reta horizontal vermelha dentro das bandas demarcadas pelas linhas tracejadas em preto. Nessas condições, o MRNULIZ tem o melhor ajuste haja visto que, os seus resíduos estão distribuídos de forma mais satisfatória dentro da região delimitada pelas bandas.

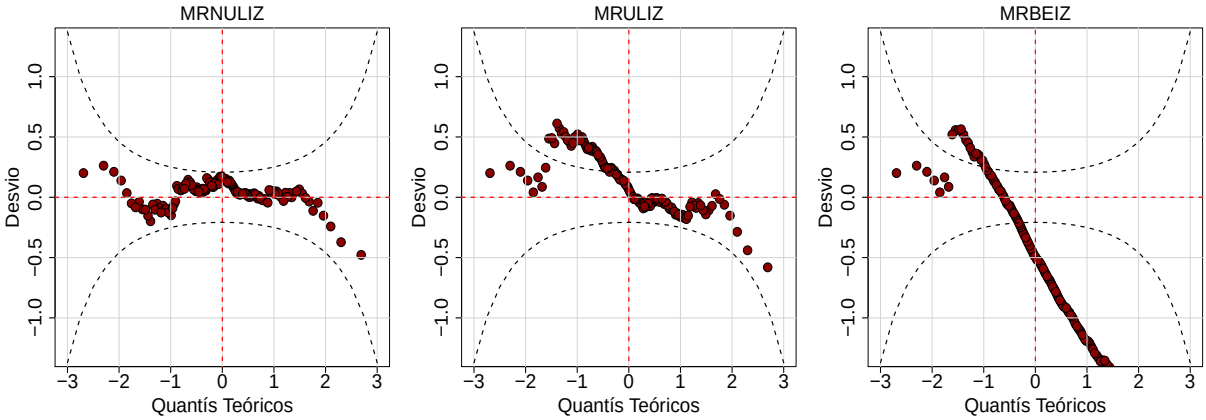

Figura 10.1 – *Worm plot* dos resíduos para modelos MRNULIZ, MRULIZ e MRBEIZ.

Tabela 3 – Resumo dos Modelos Ajustados

	MRNULIZ			MRULIZ			MRBEIZ		
	Estimativa	EP	IC de 95%	Estimativa	EP	IC de 95%	Estimativa	EP	IC de 95%
β_0	-13,841	0,886	$(-15,578\,;\,-12,104)$	-12,144	1,143	$(-14,384\,;\,-9,904)$	-7,529	1,108	$(-9,701\,;\,-5,357)$
β_1	1,216	0,126	$(0,969\,;\,1,463)$	1,472	0,176	$(1,127\,;\,1,817)$	0,310	0,169	$(-0,021\,;\,0,641)$
β_2	-0,111	0,195	$(-0,493\,;\,0,271)$	0,073	0,275	$(-0,466\,;\,0,612)$	0,002	0,267	$(-0,521\,;\,0,525)$
β_3	0,425	0,195	$(0,043\,;\,0,807)$	0,457	0,277	$(-0,086\,;\,0,100)$	0,112	0,266	$(-0,409\,;\,0,633)$
β_4	0,247	0,197	$(-0,139\,;\,0,633)$	0,561	0,281	$(0,010\,;\,1,112)$	0,110	0,268	$(-0,415\,;\,0,635)$
β_5	1,079	0,205	$(0,677\,;\,1,481)$	1,069	0,292	$(0,497\,;\,1,641)$	0,261	0,278	$(-0,284\,;\,0,806)$
β_6	-0,037	0,011	$(-0,059\,;\,-0,015)$	-0,060	0,022	$(-0,103\,;\,-0,017)$	-0,010	0,017	$(-0,043\,;\,0,023)$
β_7	4,858	1,020	$(2,859\,;\,6,857)$	2,621	1,304	$(0,065\,;\,5,177)$	0,849	1,262	$(-1,624\,;\,3,322)$
γ_0	14,390	8,191	$(-1,664\,;\,30,444)$	14,390	8,191	$(-1,664\,;\,30,444)$	14,390	8,191	$(-1,664\,;\,30,444)$
γ_1	-2,270	1,089	$(-4,404\,;\,-0,136)$	-2,270	1,089	$(-4,404\,;\,-0,136)$	-2,270	1,089	$(-4,404\,;\,-0,136)$
γ_2	-0,599	0,315	$(-1,216\,;\,0,018)$	-0,599	0,315	$(-1,216\,;\,0,018)$	-0,599	0,315	$(-1,216\,;\,0,018)$
γ_3	-0,207	0,097	$(-0,397\,;\,-0,0179)$	-0,207	0,097	$(-0,397\,;\,-0,017)$	-0,207	0,097	$(-0,397\,;\,-0,0179)$
logito(ϕ)	—	—	—	—	—	—	-2,740	0,052	$(2,638\,;\,2,842)$

11 Conclusões

Neste trabalho propomos versões inflacionadas da distribuição NUL para lidar com variáveis unitárias continuas na presença de zeros, de uns e de zeros e uns. Nossas distribuições servem de alternativas viáveis às distribuições literárias que tenham o mesmo suporte visto que contam com uma série de propriedades importantes tais como formas fechadas para os momentos e também para função quantil, são membros da família exponencial e descrevem, dentro do subintervalo $(0, 1)$, um comportamento assimétrico unimodal. O estudo de simulação mostrou que o EMV da média condicional μ sofre influencia do parâmetro de mistura σ, tendo um EQM menor para valores mais baixos de σ. Em uma aplicação a dados reais sobre a taxa de suicídio no ano de 2016, que tem por característica uma forte concentração muito próximo de zero, o MRNULIZ foi o que apresentou o melhor ajuste quando comparado ao MRULIZ e o MRBEIZ. Por fim, consideramos que nossos modelos possam ser úteis aos pesquisadores e profissionais da área que busquem por distribuições de probabilidade com características similares.

Referências

ABRAMOWITZ, M.; STEGUN, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: United States Department of Commerce, National Bureau of Standards, 1974.

BAPAT, S. R.; BHARDWAJ, R. On an Inflated Unit-Lindley Distribution. *arXiv preprint arXiv:2102.04687*, 2021.

BAPAT, S. R.; BHARDWAJ, R. On an inflated unit-lindley distribution. *arXiv preprint arXiv:2102.04687*, 2021.

CASELLA, G.; BERGER, R. L. *Statistical inference*. Califórnia: Duxbury Pacific Grove, 2002. v. 2.

CHAI, H.; JIANG, H.; LIN, L.; LIU, L. A marginalized two-part beta regression model for microbiome compositional data. *PLoS computational biology*, v. 14, n. 7, p. e1006329, 2018.

CORLESS, R. M.; GONNET, G. H.; HARE, D. E. G.; JEFFREY, D. J.; KNUTH, D. E. On the Lambert W function. *Advances in Computational Mathematics*, v. 5, n. 1, p. 329–359, 1996.

CRIBARI-NETO, F.; SANTOS, J. Inflated Kumaraswamy distributions. *Anais da Academia Brasileira de Ciências*, v. 91, n. 2, 2019.

ENEA, M.; STASINOPOULOS, M.; RIGBY, B.; HOSSAIN, A. *gamlss.inf: Fitting Mixed (Inflated and Adjusted) Distributions*. [S.l.], 2019. R package version 1.0-1. Disponível em: <https://CRAN.R-project.org/package=gamlss.inf>.

GHITANY, M. E.; MAZUCHELI, J.; MENEZES, A. F. B.; ALQALLAF, F. The unit-inverse Gaussian distribution: A new alternative to two-parameter distributions on the unit interval. *Communications in Statistics-Theory and Methods*, v. 48, n. 14, p. 3423–3438, 2019.

JODRÁ, P. Computer generation of random variables with Lindley or Poisson-Lindley distribution via the Lambert W function. *Mathematics and Computers in Simulation*, v. 81, n. 4, p. 851–859, 2010.

LINDLEY, D. V. Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 20, n. 1, p. 102–107, 1958.

LIU, P.; YUEN, K. C.; WU, L.-C.; TIAN, G.-L.; LI, T. Zero-one-inflated simplex regression models for the analysis of continuous proportion data. *Statistics and Its Interface*, v. 13, n. 2, p. 193–208, 2020.

MAZUCHELI, J.; BAPAT, S. R.; MENEZES, A. F. B. A new one-parameter unit-lindley distribution. *Chilean Journal of Statistics (ChJS)*, v. 11, n. 1, 2020.

MAZUCHELI, J.; MENEZES, A. F. B.; CHAKRABORTY, S. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, v. 46, n. 4, p. 700–714, 2019.

MAZUCHELI, J.; MENEZES, A. F. B.; DEY, S. Unit-Gompertz distribution with applications. *Statistica*, v. 79, n. 1, p. 25–43, 2019.

MAZUCHELI, J.; MENEZES, A. F. B.; GHITANY, M. E. The unit-Weibull distribution and associated inference. *Journal of Applied Probability and Statistics*, v. 13, p. 1–22, 2018.

MENEZES, A. F.; MAZUCHELI, J.; BOURGUIGNON, M. A parametric quantile regression approach for modelling zero-or-one inflated double bounded data. *Biometrical Journal*, v. 63, n. 4, p. 841–858, 2021.

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

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

QUEIROZ, F. F.; LEMONTE, A. J. A broad class of zero-or-one inflated regression models for rates and proportions. *Canadian Journal of Statistics*, v. 49, n. 2, p. 566–590, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>.

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

RIVAS, L.; CAMPOS, F. Zero inflated waring distribution. *Communications in Statistics-Simulation and Computation*, p. 1–16, 2021.

SANTOS, B.; BOLFARINE, H. Bayesian analysis for zero-or-one inflated proportion data using quantile regression. *Journal of Statistical Computation and Simulation*, v. 85, n. 17, p. 3579–3593, 2015.

SANTOS, B.; BOLFARINE, H. Bayesian quantile regression analysis for continuous data with a discrete component at zero. *Statistical Modelling*, v. 18, n. 1, p. 73–93, 2018.

SILVA, A. R.; AZEVEDO, C. L.; BAZÁN, J. L.; NOBRE, J. S. Augmented-limited regression models with an application to the study of the risk perceived using continuous scales. *Journal of Applied Statistics*, v. 48, n. 11, p. 1998–2021, 2021.

TOMARCHIO, S. D.; PUNZO, A. Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, v. 182, n. 4, p. 1247–1266, 2019.

XIE, M.; HE, B.; GOH, T. Zero-inflated poisson model in statistical process control. *Computational Statistics & data Analysis*, v. 38, n. 2, p. 191–201, 2001.

Apêndice A

```
# Implementação dos modelos no software R

rm(list = ls())
graphics.off()

library('lamW')
library('expint')
library('gamlss')
library('gamlss.inf')

# ----
# >>>> distribuição New unit-Lindley (Nul)
# ----

# if X ~ Nul(mu), then E[X] = mu.

#----
#---- função densidade de probabilidade
#----

dNul <- function(y, mu, log = FALSE){

  theta = mu/(1-mu);

  t1 = theta * theta;
  t5 = y * y;
  t12 = exp(-theta * (0.1e1 - y) / y);
  fy = t1 / (0.1e1 + theta) / t5 / y * t12;

  if(log){fy <- log(fy)}

  return(fy)
}
```

```
#----
#---- função de distribuição acumulada
#----

pNul <- function(q, mu, lower.tail = TRUE, log.p = FALSE){

  theta = mu/(1-mu);

  t4 = 0.1e1 / q;
  t6 = exp((-0.1e1 + q) * theta * t4);
  cdf = (q + theta) * t6 * t4 / (0.1e1 + theta);

  if(!lower.tail){cdf <- 1-cdf}
  if(log.p){cdf <- log(cdf)}

  return(cdf)
}

#----
#---- função quantil
#----

qNul <- function(p, mu, lower.tail = TRUE, log.p = FALSE){

  theta = mu/(1-mu);

  if(log.p){p <- exp(p)}
  if(!lower.tail){p <- 1 - p}

  t1 = 0.1e1 + theta;
  t3 = exp(-t1);
  t5 = lamW::lambertWm1(-p * t1 * t3);
  qtf = -theta/(t5 + 1);

  return(qtf)
}
```

```
#----
#---- gerar valores aleatórios
#----

rNul <- function(n, mu){

  stopifnot( n > 0)

  n <- ceiling(n);
  u <- runif(n);
  qNul(u, mu);
}



# ----------------------------
# ---------------------------- gamlss implantation
# ----------------------------

pow <- function(x, y) x^y

Nul <- function (mu.link = "logit"){
  mstats <- checklink("mu.link", "Nulindley",
                      substitute(mu.link),c("logit",
                      "probit", "cloglog", "cauchit",
                      "own"))
  structure(list(family = c("Nul", "New unit-Lindley
  distribution"),
                 parameters = list(mu = TRUE),
                 nopar = 1,
                 type = "Continuous",
                 mu.link = as.character(substitute(mu.link)
                 ),
                 mu.linkfun = mstats$linkfun,
                 mu.linkinv = mstats$linkinv,
                 mu.dr = mstats$mu.eta,
                 dldm = function(y, mu){t1 = mu * mu; t3 =
```

```
                0.2e1 * y; t10 = pow(-0.1e1 + mu, 0.2e1);
                (t1 * y + (-t3 - 0.1e1) * mu + t3) / mu /
                t10 / y },
                d2ldm2 = function(y, mu){t1 = mu * mu;
                t11 = -0.1e1 + mu; t12 = t11 * t11;
                (-y * t1 * mu + (0.3e1 * y + 0.2e1) *
                t1 - 0.6e1 * mu * y + 0.2e1 * y) /
                t11 / t12 / y / t1 },
                G.dev.incr = function(y, mu, ...){-2*
                dNul(y, mu,  log = TRUE)},
                rqres = expression(rqres(pfun = "pNul",
                type = "Continuous", y, mu)),
                mu.initial = expression(mu <- mean(y)),
                mu.valid = function(mu) all(mu>0 & mu<1),
                y.valid = function(y) all( y>0 &  y<1),
                mean = function(mu){mu},
                variance = function(mu){theta = -mu/(-1+
                mu);
                t1 = expint(theta, 1); t2 = theta * theta;
                t6 = exp(theta); return( t1 * t2 / (0.1e1
                + theta) * t6 )}),
            class = c("gamlss.family", "family"))
}

# ----
# >>>> Modelos inflacionados
# ----

# As versões inflacionadas podem ser facilmente obtidas através
# da biblioteca gamlss.inf que usa a especificação logito para
# os seus ajustes.

library('gamlss.inf')

gen.Inf0to1(family = "Nul", type.of.Inflation =  "Zero")

gen.Inf0to1(family = "Nul", type.of.Inflation = "One")
```

```
# Por outro lado, na parametrização adotada neste trabalho
# para a distribuição NULIZU, foi preciso implementar tal
# distribuição como um objeto da biblioteca gamlss. Desse modo
# temos

# -
# > A distribuição New unit-Lindley Inflacionada em Zero e Um
# -

#----
#---- função densidade de probabilidade
#----

dZOINUL <- function(y, mu, sigma, nu, log = FALSE){

  fy <- ifelse( y > 0 & y < 1, (1 - sigma)*dNUL(y, mu),
  sigma*(1-nu))
  fy <- ifelse( y==1, sigma*nu, fy)

  if(log){fy <- log(fy)}
  return(fy)
}


#----
#---- função de distribuição acumulada
#----

pZOINUL <- function(q, mu, sigma, nu, lower.tail = TRUE,
log.p = FALSE){

  cdf <- ifelse( q > 0 & q < 1, sigma*(1-nu) + (1-sigma)*
  pNUL(q = q, mu, lower.tail = TRUE, log.p = FALSE),
  sigma*(1-nu))
  cdf <- ifelse( q>=1, 1, cdf)
```

```
  if(!lower.tail){cdf <- 1-cdf}
  if(log.p){cdf <- log(cdf)}

  return(cdf)
}

#----
#---- função quantil
#----

qZOINUL <- function(p, mu, sigma, nu, lower.tail = TRUE,
log.p = FALSE){

  if(log.p){p <- exp(p)}
  if(!lower.tail){p <- 1 - p}

  p0 <- sigma*(1-nu);

  q <- ifelse( p <= sigma*(1-nu), 0, qNUL( (p - p0)/
  (1 - sigma) , mu, lower.tail = TRUE, log.p = FALSE))
  q <- ifelse( p >= 1- sigma*nu, 1, q)
  q
}

#----
#---- gerar valores aleatórios
#----

rZOINUL <- function(n, mu, sigma, nu){

  n <- ceiling(n)
  p <- runif(n)
  r <- qZOINUL(p, mu, sigma, nu = nu)
  r
}

# ----------------------------
```

```
# ---------------------------- gamlss implantation
# ----------------------------

ZOINUL <- function (mu.link ="logit", sigma.link ="logit",
nu.link ="logit"){
  mstats <- checklink("mu.link","ZOINUL",
  substitute(mu.link),
                      c("logit", "probit", "cloglog", "log"
                      ,"own"))
  dstats <- checklink("sigma.link","ZOINUL",
  substitute(sigma.link),
                      c("logit", "probit", "cloglog", "log"
                      ,"own"))
  vstats <- checklink("nu.link","ZOINUL",
  substitute(nu.link),
                      c("logit", "probit", "cloglog", "log"
                      ,"own"))
  structure(list(family = c("ZOINUL", "The Zero-and-One
  inflated New unit-Lindley distribution"),
        parameters = list(mu =TRUE, sigma =TRUE, nu =TRUE),
        nopar = 3,
        type = "Mixed",
        mu.link = as.character(substitute(mu.link)),
        sigma.link = as.character(substitute(sigma.link)),
        nu.link = as.character(substitute(nu.link)),
        mu.linkfun = mstats$linkfun,
        sigma.linkfun = dstats$linkfun,
        nu.linkfun = vstats$linkfun,
        mu.linkinv = mstats$linkinv,
        sigma.linkinv = dstats$linkinv,
        nu.linkinv = vstats$linkinv,
        mu.dr = mstats$mu.eta,
        sigma.dr = dstats$mu.eta,
        nu.dr = vstats$mu.eta,
        dldm = function(y, mu){
          t1 = mu * mu;
          t3 = 0.2e1 * y;
```

```
  t10 = pow(mu - 0.1e1, 0.2e1);
  dldm = ifelse( ((y == 0) | (y == 1)), 0,  0.1e1 /
  y / t10 / mu*(y * t1 + mu * (-t3 - 0.1e1) + t3))
  dldm
},
d2ldm2 = function(y, mu){
  t1 = mu * mu;
  t11 = mu - 0.1e1;
  t12 = t11 * t11;
  d2ldm2 = ifelse( (y == 0) | (y == 1), 0, 0.1e1 /
  t1 / y / t12 / t11 * (-t1 * mu * y + t1 * (0.3e1
  * y + 0.2e1) - 0.6e1 * mu * y + 0.2e1 * y) )
     d2ldm2
  },
  dldd = function(y, sigma){
    dldd = ifelse( ((y == 0) | (y == 1)), 0.1e1 /
    sigma, 0.1e1 / (-0.1e1 + sigma))
    dldd
  },
  d2ldd2 = function(y, sigma){
    t1 = sigma * sigma;
    t2 = pow(-0.1e1 + sigma, 0.2e1);
    d2ldd2 = ifelse( (y == 0) | (y == 1), -0.1e1 /
    t1, -0.1e1 / t2)
    d2ldd2
  },
  dldv = function(y, nu){
    dldv = ifelse( y==0,  -0.1e1 / (0.1e1 - nu), 0)
    dldv = ifelse( y==1, 0.1e1 / nu, dldv)
    dldv
  },
  d2ldv2 = function(y, nu){
    t1 = nu * nu;
    t2 = pow(0.1e1 - nu, 0.2e1);
    d2ldv2 = ifelse( y==0, -0.1e1 / t2, 0)
    d2ldv2 = ifelse( y==1,  -0.1e1 / t1, d2ldv2)
    d2ldv2
```

```
        },
        d2ldmdd = function(y){
          d2ldmdd = rep(0, length(y))
          d2ldmdd
        },
        d2ldmdv = function(y){
          d2ldmdv = rep(0, length(y))
          d2ldmdv
        },
        d2ldddv = function(y, sigma, nu){
          d2ldddv = rep(0, length(y))
          d2ldddv
        },
        G.dev.incr = function(y, mu, sigma, nu, ...){-2 *
        dZOINUL(y, mu, sigma, nu, log = TRUE)},
        rqres = expression(rqres(pfun = "pZOINUL", type =
        "Mixed",mass.p = c(0,1), prob.mp=cbind(sigma, nu)
        , y, mu, sigma, nu = nu)),
        mu.initial =expression(mu <- mean(y[y>0 & y<1])),
        sigma.initial =expression(sigma <- mean(y==0) ),
        nu.initial =expression(nu <- mean(y==1) ),
        mu.valid =function(mu) all(mu > 0 & mu < 1),
        sigma.valid =function(sigma) all(sigma>0 &
        sigma<1),
        nu.valid =function(nu) all(nu > 0 & nu < 1),
        y.valid =function(y) all(y >= 0 & y <= 1)),
    class = c("gamlss.family", "family"))
}
```