AVALIAÇÃO DOS ESTIMADORES DO MODELO DE REGRESSÃO BETA COM DISPERSÃO VARIÁVEL: UM ESTUDO DE SIMULAÇÃO

Laís Helen Loose
Universidade Federal de Santa Maria - Curso de Bacharelado em Estatística
laisloose@gmail.com

Bruna Gregory Palm
Universidade Federal de Santa Maria - Curso de Bacharelado em Estatística
brunagpalm@gmail.com

Fábio Mariano Bayer
Universidade Federal de Santa Maria - Departamento de Estatística e LACESM
bayer@ufsm.br

RESUMO

O presente trabalho avalia numericamente os estimadores pontuais e intervalares dos parâmetros do modelo de regressão beta com dispersão variável. Este modelo assume que a variável resposta possui distribuição beta com parâmetros de média e de precisão. Assume-se que o parâmetro de precisão não é constante ao longo das observações e que pode ser modelado da mesma forma que a média da variável dependente. Por meio de simulações de Monte Carlo avaliou-se as aproximações assintóticas dos estimadores de máxima verossimilhança e a taxa de cobertura dos intervalos de confiança aproximados. Os resultados numéricos confirmam a consistência dos estimadores. Também observa-se que os estimadores que modelam a precisão são percentualmente mais viesados do que os estimadores que modelam a média. Em relação aos intervalos de confiança verificam-se distorções consideráveis em pequenas amostras tanto nas inferências sobre os parâmetros do submodelo da média quanto para o submodelo da precisão.

ABSTRACT

This paper presents a numerical evaluation of the point and interval estimators in the beta regression model with varying dispersion. This model assumes that the response variable has the beta distribution with mean and precision parameters. In such models, both the mean and the dispersion depend upon independent variables. By Monte Carlo simulations we evaluated the asymptotic approximations of maximum likelihood estimators and the coverage rates of the approximated confidence intervals. The numerical results confirm the consistency of estimators. We also observe that the dispersion submodel estimators are more biased than the estimators of the mean submodel. The confidence intervals showed considerable distortions in small samples sizes.

Palavras-chave: Dispersão variável, estimadores de máxima verossimilhança, intervalos de confiança, regressão beta, simulação de Monte Carlo.

1 INTRODUÇÃO

A análise de regressão é uma das técnicas estatísticas mais utilizadas. Tem por objetivo investigar e modelar, baseada em um banco de dados, a relação entre uma variável de interesse e variáveis explicativas. Os modelos de regressão linear normal são usuais em análises empíricas. No entanto, esses modelos podem ser impróprios em situações que a variável de interesse pertence a um intervalo limitado, como taxas e proporções. Uma alternativa seria o uso de transformações da variável de interesse. Contudo, essa abordagem possui certas limitações. Além dos resultados serem interpretados em termos da média da variável transformada e não em termos da média da variável de interesse, taxas e proporções são geralmente heteroscedásticas e assimétricas, tal que modelos lineares normais podem conduzir à conclusões inferenciais distorcidas [3, 7].

Visando encontrar alternativas para situações em que os dados são restritos a um intervalo limitado, em que modelos de regressão linear normal são inapropriados, em [6] é proposto o modelo de regressão beta. Este modelo é adequado quando a variável dependente Y assume valores contínuos no intervalo $(0, 1)$, como taxas, proporções ou índices. No modelo proposto em [6] supõem-se que a variável resposta possui distribuição beta com parâmetro de precisão constante e o parâmetro de média relacionado a um preditor linear através de uma função de ligação, covariáveis e parâmetros de regressão desconhecidos [2, 13].

Para os modelos de regressão beta observa-se na literatura trabalhos que abordam melhoramentos inferenciais, análise de diagnóstico e aspectos de modelagem. Detalhes sobre inferências em grandes amostras e análise de diagnóstico nessa classe de modelos podem ser encontrados em [4, 5]. Melhoramentos em estimação pontual e intervalar são apresentados por [13]. Uma generalização do modelo de regressão beta é apresentado em [18], onde considera-se uma estrutura de regressão para o parâmetro de precisão e modelos não-lineares. Em [18] também são obtidas correções analíticas de viés para os estimadores de máxima verossimilhança, generalizando os resultados de [13]. Uma discussão a respeito de modelagem de regressão beta no sistema R [16] é apresentada com detalhes por [3]. Modelos de regressão beta inflacionados, que acomodam dados que contêm zeros e/ou uns, são tratados como extensões do modelo de regressão beta em [14]. Correções de Bartlett e Bartlett bootstrap são consideradas em [1] para o melhoramento do teste da razão de verossimilhanças em regressão beta. Em [2] são avaliados e propostos critérios de seleção para o modelo de regressão no caso de dispersão variável. Estudos de aplicações empíricas do modelo de regressão beta são apresentados em [9], [8] e [17].

A estimação dos parâmetros do modelo de regressão beta é baseada nos estimadores de máxima verossimilhança (EMV), em que os procedimentos inferenciais são semelhantes aos dos modelos lineares generalizados [11]. Esses estimadores possuem propriedades de consistência e normalidade assintótica, ou seja, sendo $\hat{\theta}$ o EMV de um parâmetro θ, para n suficientemente grande, $\hat{\theta}$ tem distribuição aproximadamente normal com média θ e variância igual ao inverso da informação de Fisher ($I_F(\theta)$). As inferências são feitas baseadas nessa aproximação assintótica que em pequenas amostras pode conduzir a resultados inferenciais distorcidos tanto em termos de viés do estimador pontual, quanto em termos de taxa de cobertura do intervalo de confiança. Neste sentido, torna-se importante a avaliação numérica dessas distorções em amostras de tamanho finito, sendo o objetivo do presente trabalho.

Este artigo está organizado da seguinte forma. A Seção 2 introduz o modelo de regressão beta com dispersão variável. São apresentadas a função de log-verossimilhança, a função escore e a matriz de informação de Fisher. Na Seção 3 é descrito o experimento de simulação de Monte Carlo. Na Seção 4 são apresentados os resultados da simulação, bem como a sua discussão. Por fim, na Seção 5 estão apresentadas as conclusões do trabalho.

2 O MODELO DE REGRESSÃO BETA COM DISPERSÃO VARIÁVEL

A densidade beta é dada por:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, 0 < y < 1, \tag{1}$$

em que $p, q > 0$ e $\Gamma(\cdot)$ é a função gama, isto é $\Gamma(u) = \int_0^\infty t^{u-1}e^{-t}\mathrm{dt}$.

Com essa parametrização, a média e a variância de uma variável aleatória Y que possui distribuição beta são dadas, respectivamente, por:

$$E(Y) = \frac{p}{(p+q)},$$
$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

O modelo de regressão beta proposto por [6] utiliza uma reparametrizção da densidade beta dada em (1). A parametrização proposta é indexada pelos parâmetros de média e de precisão, da seguinte forma: $\mu = p/(p+q)$ e $\phi = p+q$, consequentemente $p = \mu\phi$ e $q = (1-\mu)\phi$. Com isso:

$$E(Y) = \mu,$$
$$Var(Y) = \mu(1-\mu)/(1+\phi),$$

em que μ é a média de Y e ϕ é o parâmetro de precisão, que é o inverso da dispersão. Deste modo, a densidade (1) pode ser escrita como

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

em que $0 < \mu < 1$ e $\phi > 0$.

O modelo proposto por [6] considera ϕ constante, ou seja, a dispersão é considerada fixa ao longo das observações. Contudo, ao supor ϕ constante erroneamente as perdas de eficiência podem ser substanciais. Fato que pode ser visualizado através da Figura 1, que apresenta as densidades estimadas dos estimadores do parâmetro de inclinação $\beta_1 = 1$, com e sem modelagem da dispersão. As estimativas das densidades foram obtidas de uma simulação de Monte Carlo com 10000 réplicas. O processo gerador dos dados incluía dispersão variável em um modelo de regressão beta com a média dada por $\mathrm{logit}(\mu_t) = \beta_0 + \beta_1 \mathrm{x_{t1}} + \beta_2 \mathrm{x_{t2}}$, e modelo para precisão por $\log(\phi_t) = \gamma_0 + \gamma_1 z_{t1} + \gamma_2 z_{t2}$. Analisando a Figura 1, constata-se que a estimação eficiente (menor variância) dos parâmetros da regressão depende da modelagem correta da dispersão.

O modelo de regressão beta com dispersão variável, já discutido em [18] e [7], pode ser entendido como uma extensão do modelo proposto por [6]. O parâmetro de precisão não é constante ao longo das observações, sendo modelado em termos de covariáveis e de parâmetros desconhecidos, através de uma estrutura de regressão, da mesma forma que a média. Como visto na Figura 1, essa abordagem resulta em estimadores mais eficiente dos parâmetros de regressão.

Seja $Y_1, \ldots, Y_n$ variáveis aleatórias independentes, em que cada Y_t, $t = 1, \ldots, n$, tem distribuição dada por (2), ou seja, $Y_t \sim \mathrm{Beta}(\mu_t, \phi_t)$. Para o modelo de regressão beta com dispersão variável, encontra-se a seguinte relação:

$$g_1(\mu_t) = \sum_{i=1}^{r} x_{ti}\beta_i = \eta_{1t},$$
$$g_2(\phi_t) = \sum_{i=1}^{s} z_{ti}\gamma_i = \eta_{2t},$$

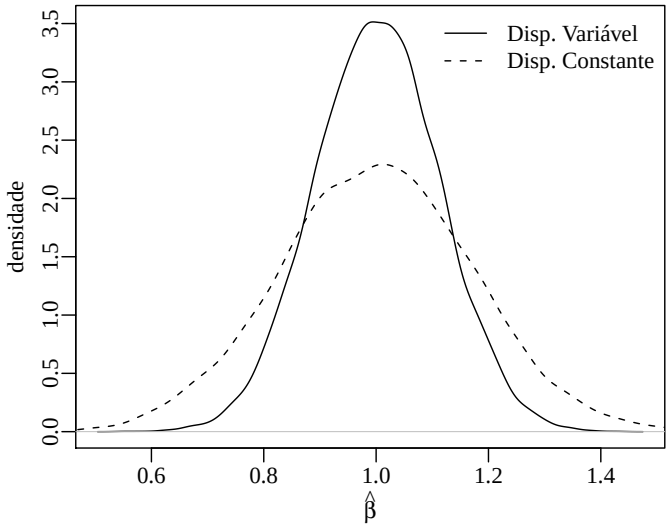


FIGURA 1: Densidades estimadas dos estimadores de β_1, com e sem modelagem da dispersão.

em que $\beta = (\beta_1, \ldots, \beta_r)^\top \in \mathbb{R}^r$ e $\gamma = (\gamma_1, \ldots, \gamma_s)^\top \in \mathbb{R}^s$ são os vetores de parâmetros desconhecidos a serem estimados para a média e para a precisão, respectivamente, $r+s = k < n$, $\eta_{1t} = x_t^T\beta$ e $\eta_{2t} = z_t^T\gamma$ são os preditores lineares, $x_t^T = (x_{t1}, \ldots, x_{tr})$ e $z_t^T = (z_{t1}, \ldots, z_{ts})$ representam as variáveis explicativas assumidas fixas e conhecidas, $g_1(\cdot)$ e $g_2(\cdot)$ são as funções de ligação estritamente monótonas e duplamente diferenciáveis, tais que $g_1(0,1) \to \mathbb{R}$ e $g_2(0,\infty) \to \mathbb{R}$ [2, 12, 18].

Diferentes funções de ligação podem ser utilizadas nos modelos de regressão beta. Para $g_1(\cdot)$ as usuais são a logit, $g_1(\mu) = \log[\mu/(1-\mu)]$, a probit, $g_1(\mu) = \Phi^{-1}(\mu)$, em que $\Phi(\cdot)$ é a função de distribuição normal, a complemento log-log, $g_1(\mu) = \log[-\log(1-\mu)]$, entre outras. Para $g_2(\cdot)$ as usuais são a logarítmica, $g_2(\phi) = \log(\phi)$ e a função raiz quadrada $g_2(\phi) = \sqrt{\phi}$ [7, 18].

Para a estimação dos vetores paramétricos β e γ são utilizados os estimadores de máxima verossimilhança. A partir de uma amostra de n observações o logaritmo da função de verossimilhança é dado por:

$$\ell(\beta,\gamma) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t), \tag{3}$$

em que

$$\ell_t(\mu_t,\phi_t) = \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) + (\mu_t\phi_t - 1)\log y_t + ((1-\mu_t)\phi_t - 1)\log(1-y_t)],$$

$\mu_t = g_1^{-1}(\eta_{1t})$ e $\phi_t = g_2^{-1}(\eta_{2t})$ são funções de β e γ, respectivamente. Ao derivar a função de log-verossimilhança dada em (3) em relação a β_R, com $R = 1, \ldots, r$, a função escore é dada por:

$$U_R(\beta,\gamma) = \frac{\partial\ell(\beta,\gamma)}{\partial\beta_R} = \sum_{t=1}^{n} \phi_t(y_t^* - \mu_t^*)\frac{d\mu_t}{d\eta_{1t}}\frac{\partial\eta_{1t}}{\partial\beta_R},$$

em que, $y_t^* = \log(y_t/(1-y_t))$, $\mu_t^* = \psi(\mu_t\phi_t) - \psi((1-\mu_t)\phi_t)$, sendo $\psi(\cdot)$ a função digamma, isto é $\psi(u) = \frac{d\log\Gamma(u)}{du}$, para $u > 0$. Derivando em relação a γ_S, com $S = 1, \ldots, s$, tem-se:

$$U_S(\beta,\gamma) = \frac{\partial\ell(\beta,\gamma)}{\partial\gamma_S} = \sum_{t=1}^{n} \mu_t(y_t^* - \mu_t^*) + \psi(\phi_t) - \psi((1-\mu_t)\phi_t) + \log(1-y_t)\frac{d\phi_t}{d\eta_{2t}}\frac{\partial\eta_{2t}}{\partial\gamma_S}.$$

O vetor escore relativo a β é dado por:

$$U_\beta(\beta,\gamma) = X^\top \Phi \mathrm{T}_1(\mathrm{y}^* - \mu^*),$$

em que X é uma matriz $n \times r$ cuja t-ésima linha é x_t, $\mathrm{T}_1 = \mathrm{diag}\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\eta_{11}}, \ldots, \frac{\mathrm{d}\mu_\mathrm{n}}{\mathrm{d}\eta_{1\mathrm{n}}}\right)$, $\Phi = \mathrm{diag}(\phi_1, \ldots, \phi_\mathrm{n})$.

O vetor escore relativo a γ é dado por:

$$U_\gamma(\beta,\gamma) = Z^\top \mathrm{T}_2 v,$$

em que Z é uma matriz $n \times s$ cuja t-ésima linha é z_t, $\mathrm{T}_2 = \mathrm{diag}\left(\frac{\mathrm{d}\phi_1}{\mathrm{d}\eta_{21}}, \ldots, \frac{\mathrm{d}\phi_\mathrm{n}}{\mathrm{d}\eta_{2\mathrm{n}}}\right)$, $v = (v_1, \ldots, v_n)$, sendo $v_t = \mu_t(y_t^* - \mu_t) + \psi(\phi_t) - \psi((1-\mu_t)\phi_t) + \log(1-y_t)$.

Os estimadores de máxima verossimilhança para o modelo de regressão beta são obtidos a partir da resolução do seguinte sistema:

$$\begin{cases} U_\beta(\beta,\gamma) = & 0 \\ U_\gamma(\beta,\gamma) = & 0 \end{cases}.$$

Para a solução deste sistema não se verifica uma forma fechada, sendo necessário o uso de algoritmos de otimização não-linear para encontrar as estimativas de máxima verossimilhança. Usualmente, utiliza-se o método quasi-Newton BFGS [15].

A matriz de informação de Fischer conjunta para β e γ, é dada por:

$$I_F(\beta,\gamma) = \begin{pmatrix} I_{(\beta,\beta)} I_{(\beta,\gamma)} \\ I_{(\gamma,\beta)} I_{(\gamma,\gamma)} \end{pmatrix},$$

em que $I_{(\beta,\beta)} = X^\top \Phi T V^* T \Phi X$, $I_{(\beta,\gamma)} = (I_{(\gamma,\beta)})^\top = X^\top \Phi T (MV^* + C) H Z$ e $I_{(\gamma,\gamma)} = Z^\top H (M^2 V^* + 2MC + V^\dagger)$. Em que $M = \mathrm{diag}(\mu_1, \ldots, \mu_\mathrm{n})$, $C = \mathrm{diag}(\mathrm{c}_1, \ldots, \mathrm{c}_\mathrm{n})$ e $T = \mathrm{diag}\left(\frac{1}{\mathrm{g}_1'(\mu_1)}, \ldots, \frac{1}{\mathrm{g}_1'(\mu_\mathrm{n})}\right)$, $H = \mathrm{diag}\left(\frac{1}{\mathrm{g}_2'(\phi_1)}, \ldots, \frac{1}{\mathrm{g}_2'(\phi_\mathrm{n})}\right)$, $V^* = \mathrm{diag}(\mathrm{v}_1^*, \ldots, \mathrm{v}_\mathrm{n}^*)$ e $V^\dagger = \mathrm{diag}(\mathrm{v}_1^\dagger, \ldots, \mathrm{v}_\mathrm{n}^\dagger)$. Em que: $c_t = -\psi'((1-\mu_t)\phi_t)$, $v_t^* = \psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t)$ e $v_t^\dagger = \psi'((1-\mu_t)\phi_t) - \psi'(\phi_t)$.

Sob certas condições de regularidade, para tamanhos amostrais grandes, a distribuição conjunta de $\widehat{\beta}$ e $\widehat{\gamma}$ é aproximadamente normal $k-$multivariada, dada por:

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\gamma} \end{pmatrix} \sim N_k \left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, I_F(\beta,\gamma)^{-1} \right),$$

em que $\widehat{\beta}$ e $\widehat{\gamma}$ são os estimadores de máxima verossimilhança de β e γ, respectivamente.

3 SIMULAÇÃO DE MONTE CARLO

A avaliação das inferências pontuais e intervalares no modelo de regressão beta foi realizada por meio de simulações de Monte Carlo. A implementação computacional foi desenvolvida em linguagem `R` [16] sendo que para a estimação dos parâmetros foi utilizada a função `betareg` [3]. Os resultados numéricos apresentados na Seção 4 são baseados no modelo de regressão beta com as estruturas da média e do parâmetro de precisão, respectivamente, dadas por:

$$g_1(\mu_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2},$$
$$g_2(\phi_t) = \gamma_0 + \gamma_1 z_{t1} + \gamma_2 z_{t2},$$

em que $t = 1, \ldots, n$. Para a estrutura de regressão da média $g_1(\mu_t)$, foi utilizada a função de ligação logit e para a estrutura do parâmetro de precisão $g_2(\phi_t)$ a função de ligação logarítmica. Dessa forma, μ_t e ϕ_t são dados, respectivamente, por:

$$\mu_t = \exp(\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2})/(1 + \exp(\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2}), \tag{4}$$

$$\phi_t = \exp(\gamma_0 + \gamma_1 z_{t1} + \gamma_2 z_{t2}). \tag{5}$$

Na simulação de Monte Carlo foram fixados $\beta_0 = 0,5$, $\beta_1 = -0,5$, $\beta_2 = -0,5$, investigando diferentes cenários para ϕ. Estes cenários implicam em valores médios de ϕ em torno de 20 (grande dispersão), 100 (dispersão intermediária) e 200 (pequena dispersão). Sendo os valores de γ_i dados, respectivamente, para esses três cenários, por: (i) $\gamma_0 = 2$, $\gamma_1 = 1$, $\gamma_2 = 1$, (ii) $\gamma_0 = 2,3$, $\gamma_1 = 2,2$, $\gamma_2 = 2$ e (iii) $\gamma_0 = 3$, $\gamma_1 = 2,5$, $\gamma_2 = 1,5$. A matriz de regressores é gerada a partir de uma distribuição uniforme padrão, $\mathcal{U}(0,1)$ e permanece constante durante todas as réplicas de Monte Carlo. O número de réplicas foi de 10000, sendo utilizados tamanhos amostrais iguais a $n = 20, 40, 70, 150$. Para cada réplica gera-se uma amostra $y_1, \ldots, y_n$ com densidade beta dada por (2), em que μ_t e ϕ_t são dadas por (4) e (5).

A fim de avaliar numericamente os estimadores pontuais se faz necessária a utilização de algumas medidas. Baseadas nas 10000 réplicas dos estimadores de máxima verossimilhança dos parâmetros do modelo foram calculados a média, o viés percentual relativo (VR%) definido como $\{E(\widehat{\theta}) - \theta\}/\theta$, a variância (Var), o erro quadrático médio (EQM), o coeficiente de assimetria (CA) e o coeficiente de curtose (K).

Para a avaliação da estimação intervalar são calculadas taxas de cobertura (TC), sendo utilizada significância $\alpha = 0,05$. Em cada réplica calcula-se um intervalo de confiança para os parâmetros e verifica-se se o parâmetro pertence ou não ao intervalo. A taxa de cobertura é dada pela porcentagem de réplicas em que o parâmetro pertence de fato ao intervalo de confiança [10, 12]. O esperado é que os valores da taxa de cobertura se aproximem do valor do coeficiente de confiança $(1 - \alpha)$, ou seja, espera-se que a taxa de cobertura esteja próxima ao valor $0,95$.

4 RESULTADOS NUMÉRICOS

As Tabelas 1, 2 e 3 apresentam os resultados da avaliação numérica dos estimadores dos parâmetros β_i e γ_i, $i = 0, 1, 2$, para diferentes tamanhos amostrais.

Ao analisar a Tabela 1 espera-se que o viés relativo se encontre próximo de zero. Ao considerar as estimativas de Monte Carlo dos parâmetros de regressão verifica-se que o viés relativo dos $\widehat{\beta}_i$ são próximos de zero, independente do tamanho amostral. Já o estimador $\widehat{\gamma}$ apresenta viés relativo considerável, mas que diminui a medida que a amostra aumenta.

Em relação ao erro quadrático médio, verifica-se que diminui à medida que o tamanho amostral aumenta. Fato que indica consistência dos estimadores de máxima verossimilhança. Ao considerar a propriedade de normalidade assintótica, espera-se que a distribuição dos estimadores do modelo se aproxime, a medida que n aumenta, da distribuição normal. Baseado nessa propriedade, espera-se que o coeficiente de assimetria se aproxime de zero e o coeficiente de curtose de três. A aproximação normal para os estimadores de β_i, nas amostras menores, se mostra mais adequada do que assumir normalidade para $\widehat{\gamma}_i$. Para $n = 20$ e $\widehat{\gamma}_0$, por exemplo, tem-se valores de assimetria e curtose iguais a $0,59$ e $4,36$, respectivamente. No entanto, à medida que aumenta o tamanho amostral esses valores se tornam próximos de zero e três.

Baseados na distribuição assintótica dos EMV considera-se intervalos de confiança usuais com coeficiente de confiança de $0,95$. Espera-se que a taxa de cobertura esteja próxima desse valor. Para tamanhos amostrais pequenos as taxas de cobertura são bem inferiores ao valor de $0,95$. Para $n = 20$, por exemplo, a TC se mostra em torno de $0,78$. Verifica-se que à medida que o tamanho amostral aumenta, o valor da taxa de cobertura se torna mais próximo do coeficiente de confiança $0,95$.

TABELA 1: Resultados da simulação de Monte Carlo considerando o modelo de regressão beta com dispersão variável utilizando $\beta_0 = 0,5$, $\beta_1 = -0,5$, $\beta_2 = -0,5$, $\gamma_0 = 2$, $\gamma_1 = 1$ e $\gamma_2 = 1$, (grande dispersão).

Tamanho amostral = 20							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	−0,32	0,12	0,12	−0,02	3,35	0,79
$\widehat{\beta_1}$	−0,50	0,35	0,25	0,25	0,02	3,29	0,79
$\widehat{\beta_2}$	−0,50	−0,66	0,20	0,20	−0,04	3,33	0,78
$\widehat{\gamma_0}$	2,18	9,25	2,34	2,38	0,59	4,36	0,79
$\widehat{\gamma_1}$	1,21	21,22	4,75	4,82	0,14	3,52	0,75
$\widehat{\gamma_2}$	1,18	18,52	5,90	5,95	0,07	3,80	0,74
Tamanho amostral = 40							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	0,53	0,05	0,05	0,01	2,98	0,89
$\widehat{\beta_1}$	−0,50	0,08	0,08	0,07	0,00	2,96	0,90
$\widehat{\beta_2}$	−0,50	0,67	0,08	0,07	−0,01	3,06	0,89
$\widehat{\gamma_0}$	2,09	4,40	0,45	0,46	0,38	3,35	0,89
$\widehat{\gamma_1}$	1,10	9,84	0,85	0,85	0,04	3,38	0,88
$\widehat{\gamma_2}$	1,07	7,03	1,06	1,07	−0,11	3,16	0,87
Tamanho amostral = 70							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	0,51	0,02	0,02	0,06	3,01	0,92
$\widehat{\beta_1}$	−0,50	−0,04	0,03	0,03	−0,05	3,02	0,92
$\widehat{\beta_2}$	−0,50	0,81	0,03	0,03	−0,00	3,03	0,92
$\widehat{\gamma_0}$	2,05	2,65	0,21	0,21	0,26	3,24	0,92
$\widehat{\gamma_1}$	1,01	1,52	0,46	0,47	−0,10	3,19	0,91
$\widehat{\gamma_2}$	1,07	7,14	0,40	0,40	0,08	3,16	0,92
Tamanho amostral = 150							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	0,20	0,01	0,01	0,01	2,98	0,94
$\widehat{\beta_1}$	−0,50	0,34	0,02	0,02	−0,02	2,98	0,94
$\widehat{\beta_2}$	−0,50	0,15	0,01	0,02	0,01	3,03	0,94
$\widehat{\gamma_0}$	2,02	1,09	0,07	0,08	0,16	3,13	0,93
$\widehat{\gamma_1}$	1,02	1,69	0,16	0,16	−0,02	3,15	0,94
$\widehat{\gamma_2}$	1,02	2,28	0,17	0,17	−0,05	3,09	0,94

Ao analisar a Tabela 2, percebe-se que as médias dos $\widehat{\beta_i}$ são iguais aos valores que foram propostos aos parâmetros, já para os $\widehat{\gamma_i}$, à medida que aumenta o tamanho amostral, os valores das médias ficam mais próximos dos valores esperados. O viés raltivo para os $\widehat{\beta_i}$ não apresenta variações, se mantendo próximo de zero. Por outro lado, o viés dos $\widehat{\gamma_i}$ são consideráveis em pequenas amostras, alcançando viés relativo de 23% para $n = 20$. Contudo, com o aumento do tamanho amostral o viés relativo de $\widehat{\gamma_i}$ diminui.

Observa-se ainda na Tabela 2 a propriedade de consistência dos estimadores de máxima verossimilhança, considerando o fato do erro quadrático médio diminuir a medida que o tamanho amostral aumenta. Os valores dos coeficientes de assimetria e curtose indicam que os estimadores seguem distribuição normal assintótica, pois o coeficiente de assimetria se aproxima de zero e o de curtose de três quando aumenta o tamanho amostral. Ao analisar a taxa de cobertura, a medida que o tamanho amostral aumenta, seu valor fica mais próximo do ideal $0,95$.

Considerando a Tabela 3, os resultados são semelhantes às Tabelas 1 e 2. Percebe-se que os valores da média de $\widehat{\beta_i}$ não apresentam diferenças em relação ao valores reais dos parâmetros. Sendo que os valores dos $\widehat{\gamma_i}$ apresentam uma diferença considerável. Ao analisar o viés relativo verifica-se que para os $\widehat{\gamma_i}$, este diminui a medida que a amostra aumenta, sendo que os valores para $\widehat{\beta_i}$ apresentaram-se constantes.

Ao analisar o erro quadrático médio verifica-se que este diminui a medida que aumenta

TABELA 2: Resultados da simulação de Monte Carlo considerando o modelo de regressão beta com dispersão variável utilizando $\beta_0 = 0,5$, $\beta_1 = -0,5$, $\beta_2 = -0,5$, $\gamma_0 = 2,3$, $\gamma_1 = 2,2$, $\gamma_2 = 2$, (dispersão intermediária).

Tamanho amostral = 20							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	−0,25	0,03	0,03	0,08	3,25	0,77
$\widehat{\beta_1}$	−0,50	0,06	0,05	0,05	0,04	3,77	0,78
$\widehat{\beta_2}$	−0,50	−0,24	0,05	0,05	−0,08	3,27	0,76
$\widehat{\gamma_0}$	2,25	−1,94	2,34	2,36	0,57	4,91	0,80
$\widehat{\gamma_1}$	2,71	23,00	5,29	5,53	0,07	4,09	0,73
$\widehat{\gamma_2}$	2,36	17,87	6,05	6,20	−0,01	3,67	0,74
Tamanho amostral = 40							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	−0,31	0,01	0,01	−0,00	3,10	0,90
$\widehat{\beta_1}$	−0,50	−0,36	0,02	0,02	0,04	2,99	0,90
$\widehat{\beta_2}$	−0,50	−0,27	0,01	0,01	−0,04	3,13	0,89
$\widehat{\gamma_0}$	2,32	0,86	0,59	0,59	0,29	3,33	0,90
$\widehat{\gamma_1}$	2,40	8,85	1,04	1,09	−0,01	3,27	0,88
$\widehat{\gamma_2}$	2,12	6,07	1,09	1,09	−0,12	3,24	0,88
Tamanho amostral = 70							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	0,03	0,00	0,00	−0,01	2,97	0,92
$\widehat{\beta_1}$	−0,50	0,09	0,01	0,01	−0,01	2,95	0,92
$\widehat{\beta_2}$	−0,50	−0,13	0,01	0,01	−0,02	2,91	0,93
$\widehat{\gamma_0}$	2,33	1,10	0,22	0,22	0,20	3,11	0,93
$\widehat{\gamma_1}$	2,25	2,50	0,50	0,50	−0,09	3,05	0,92
$\widehat{\gamma_2}$	2,07	3,37	0,42	0,43	−0,03	3,15	0,92
Tamanho amostral = 150							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	0,08	0,00	0,00	−0,01	2,97	0,94
$\widehat{\beta_1}$	−0,50	0,17	0,00	0,00	0,05	3,01	0,94
$\widehat{\beta_2}$	−0,50	−0,01	0,00	0,00	−0,02	2,92	0,94
$\widehat{\gamma_0}$	2,30	−0,15	0,11	0,11	0,14	3,04	0,94
$\widehat{\gamma_1}$	2,25	2,06	0,18	0,18	−0,03	3,10	0,93
$\widehat{\gamma_2}$	2,04	1,91	0,17	0,17	−0,02	3,16	0,94

o tamanho amostral, principalmente para $\widehat{\gamma_i}$ que apresenta valores maiores de EQM em amostras menores. Considerando os coeficientes de assimetria e curtose tem-se que o coeficiente de assimetria se aproxima de zero e o de curtose de três quando aumenta o tamanho amostral.

Em relação a taxa de cobertura, observa-se que para os tamanhos amostrais menores os valores são muito inferiores a $0,95$. No entanto, a medida que aumenta o tamanho amostral, seu valor fica mais próximo do coeficiente de confiança.

A Figura 2 apresenta um resumo gráfico do viés relativo dos estimadores. Pode-se observar que o viés relativo do estimador de β_i encontra-se próximo de zero nos diferentes cenários de níveis de dispersão. Observa-se ainda, nas Figuras 2(b) e 2(c), que $\widehat{\gamma_0}$ apresentou valores de viés relativo próximos de zero. Já os estimadores de γ_1 e γ_2, em tamanhos amostrais pequenos, apresentam viés relativo com valores consideravelmente diferentes de zero. Nota-se, no entanto, que a medida que a amostra aumenta os valores de viés relativo decrescem, tornando-se próximos a zero quando $n = 150$.

Considerando as Tabelas 1, 2 e 3, pode-se observar que ao aumentar a precisão (diminuir a dispersão), ou seja, quando os valores de $\widehat{\gamma_i}$ são maiores, Tabelas 2 e 3, os vieses dos parâmetros $\widehat{\beta_i}$ são menores em amostras grandes. O EQM de $\widehat{\beta_i}$ apresenta valores menores quando temos $\widehat{\gamma_i}$ maior.

Os estimadores pontuais de β_i se mostram menos viesados do que os de γ_i, ou seja, os

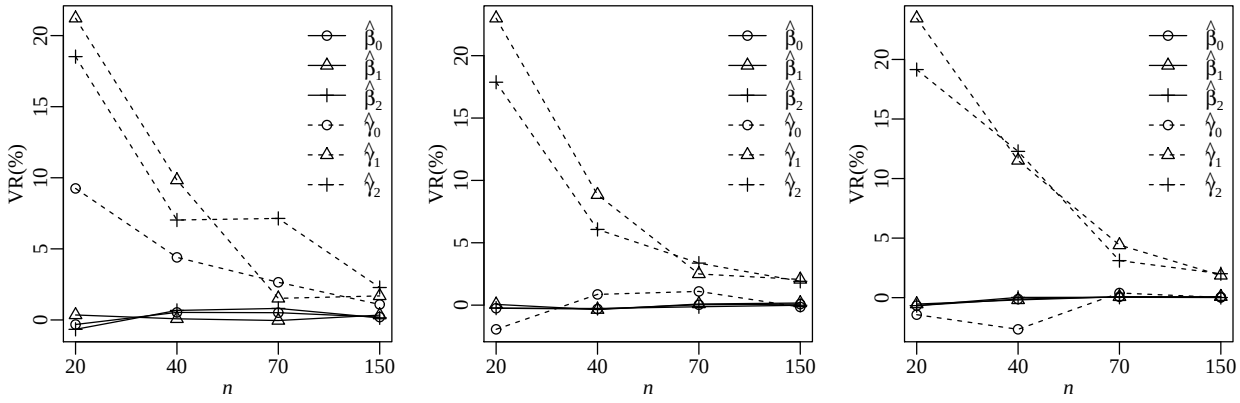
TABELA 3: Resultados da simulação de Monte Carlo considerando o modelo de regressão beta com dispersão variável utilizando $\beta_0 = 0,5$, $\beta_1 = -0,5$, $\beta_2 = -0,5$, $\gamma_0 = 3$, $\gamma_1 = 2,5$ e $\gamma_2 = 1,5$, (pequena dispersão).

Tamanho amostral = 20							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	−0,67	0,02	0,02	0,00	3,25	0,78
$\widehat{\beta_1}$	−0,50	−0,54	0,03	0,03	0,02	3,48	0,78
$\widehat{\beta_2}$	−0,50	−0,67	0,03	0,03	0,01	3,18	0,76
$\widehat{\gamma_0}$	2,96	−1,43	2,43	2,44	0,35	4,43	0,79
$\widehat{\gamma_1}$	3,09	23,46	5,11	5,44	0,06	4,11	0,74
$\widehat{\gamma_2}$	1,79	19,15	6,3	6,37	0,13	3,86	0,74
Tamanho amostral = 40							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	−0,12	0,00	0,00	0,02	3,02	0,88
$\widehat{\beta_1}$	−0,50	−0,19	0,01	0,01	−0,01	3,02	0,88
$\widehat{\beta_2}$	−0,50	0,02	0,07	0,00	−0,03	2,96	0,88
$\widehat{\gamma_0}$	2,92	−2,66	0,67	0,68	0,15	3,25	0,87
$\widehat{\gamma_1}$	2,79	11,52	0,90	0,99	−0,02	3,13	0,87
$\widehat{\gamma_2}$	1,68	12,28	1,12	1,16	−0,10	3,42	0,86
Tamanho amostral = 70							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	0,07	0,00	0,00	−0,03	3,01	0,92
$\widehat{\beta_1}$	−0,50	0,07	0,00	0,00	0,04	3,02	0,92
$\widehat{\beta_2}$	−0,50	0,03	0,00	0,00	−0,02	3,01	0,92
$\widehat{\gamma_0}$	3,01	0,39	0,21	0,21	0,16	3,20	0,92
$\widehat{\gamma_1}$	2,61	4,42	0,45	0,46	0,05	3,17	0,91
$\widehat{\gamma_2}$	1,55	3,11	0,41	0,40	−0,07	3,19	0,92
Tamanho amostral = 150							
Estimador	Média	VR%	Var	EQM	CA	K	TC
$\widehat{\beta_0}$	0,50	0,04	0,00	0,00	−0,01	2,97	0,94
$\widehat{\beta_1}$	−0,50	0,08	0,00	0,00	0,01	2,99	0,94
$\widehat{\beta_2}$	−0,50	0,01	0,00	0,00	0,00	2,99	0,94
$\widehat{\gamma_0}$	3,00	0,02	0,10	0,10	0,16	3,10	0,94
$\widehat{\gamma_1}$	2,55	1,87	0,17	0,17	0,02	3,06	0,93
$\widehat{\gamma_2}$	1,53	2,01	0,19	0,19	0,03	3,09	0,94

estimadores dos parâmetros que modelam a precisão mostram-se viesados em pequenas amostras, sendo que o viés diminui a medida que aumenta o tamanho amostral. Também percebe-se que os estimadores de γ_i possuem maior variabilidade do que os estimadores de β_i, comparando as variâncias das Tabelas 1, 2 e 3. Esses resultados corroboram com os apresentados em [18], em que são propostas correções analíticas de segunda ordem para os estimadores dos parâmetros do modelo de regressão beta com modelagem da dispersão.

Em termos de estimação intervalar, assim como os resultados apresentados em [13], para o modelo de regressão beta com dispersão constante, e em [10], para o modelo logístico hierárquico de dois níveis, os intervalos de confiança aproximados se mostram distorcidos em pequenas amostras. Contudo, a medida que o tamanho amostral aumenta as taxas de cobertura se tornam mais próximas do nível de confiança dos intervalos.

Verifica-se que, ao contrário dos estimadores pontuais, os intervalos de confiança se mostram distorcidos em pequenas amostras tanto nas inferências sobre os parâmetros do modelo da média quanto sobre os parâmetros da estrutura da precisão. Essas distorções em pequenas amostras indicam que a aproximação assintótica de normalidade dos estimadores é pobre nesses tamanhos amostrais e que intervalos de confiança corrigidos para o modelo de regressão beta com dispersão variável devem ser considerados em pesquisas futuras.



(a) Grande dispersão (b) Dispersão intermediária (c) Pequena dispersão

FIGURA 2: Viés relativo (VR) dos estimadores dos parâmetros β_i e γ_i, $i = 0, 1, 2$, em relação ao tamanho amostral, para os três cenários de níveis de dispersão.

5 CONCLUSÕES

Considerando simulações de Monte Carlo observou-se que os estimadores de máxima verossimilhança dos parâmetros que modelam a média (β_i) no modelo de regressão beta possuem boas propriedades. Verificou-se que, mesmo em amostras pequenas, o viés relativo dos estimadores dos parâmetros β_i são pequenos e que o EQM converge rapidamente à zero quando o tamanho amostral aumenta. Já os estimadores dos parâmetros γ_i são consideravelmente viesados. O viés relativo diminui em amostras maiores, mas mostra-se com um decaimento lento para zero. Os estimadores intervalares são satisfatórios em tamanhos amostrais grandes, tanto para β_i quanto para γ_i, possuindo taxas de cobertura próximas ao nível de confiança de 95%. No entanto, em pequenas amostras os intervalos de confiança possuem grandes distorções, indicando que a aproximação assintótica pode ser pobre nestes casos.

A presença de viés nos estimadores, principalmente para o submodelo da precisão, e as distorções dos intervalos de confiança indicam a necessidade de se considerar estimadores corrigidos. Em trabalhos futuros serão utilizados métodos bootstrap para melhorar essas inferências em pequenas amostras.

AGRADECIMENTOS

Os autores agradecem ao Programa IC-REUNI/UFSM e à FAPERGS pelo auxílio financeiro recebido.

REFERÊNCIAS

[1] F. Bayer e F. Cribari-Neto: *Bartlett corrections in beta regression models.* Journal of Statistical Planning and Inference, 143(3):531–547, 2013.

[2] F. M. Bayer: *Modelagem e Inferência em regressão beta.* Tese de Doutoramento, Universidade Federal de Pernambuco, 2011.

[3] F. Cribari-Neto e A. Zeileis: *Beta Regression in R.* Journal of Statistical Software, 34(2), 2010.

[4] P. Espinheira, S. L. P. Ferrari e F. Cribari-Neto: *On beta regression residuals.* Journal of Applied Statistics, 35:407–419, 2008.

[5] P. L. Espinheira, S. L. P. Ferrari e F. Cribari-Neto: *Influence Diagnostics in Beta Regression*. Computational Statistics & Data Analysis, 52:4417–4431, 2008.

[6] S. L. P. Ferrari e F. Cribari-Neto: *Beta regression for modelling rates and proportions*. Journal of Applied Statistics, 31(7):799–815, 2004.

[7] S. L. P. Ferrari e E. C. Pinheiro: *Improved likelihood inference in beta regression*. Journal of Statistical Computation and Simulation, 81(4):431–443, 2011.

[8] D. Hancox, C. J. Hoskin e R. S. Wilson: *Evening up the score: Sexual selection favours both alternatives in the colour-polymorphic ornate rainbowfish*. Animal Behaviour, 80:845–851, 2010.

[9] Y. Kajita, E. O'Neill, Y. Zheng, J. Obrycki e D. Weisrock: *A population genetic signature of human releases in an invasive ladybeetle*. Molecular Ecology, 21:5473–5483, 2012.

[10] J. Lemonte e T. F. N. M. Silva: *Estimação pontual e intervalar no modelo logístico hierárquico de dois níveis*. Revista de Matemática e Estatística, 24(2), 2006.

[11] P. McCullagh e J. Nelder: *Generalized linear models*. Chapman and Hall, 2nd ed., 1989.

[12] M. S. Oliveira e S. L. P. Ferrari: *Inferência em um modelo de Regressão Beta: resultados numéricos*. Em *49 Reunião da RBRAS*, 2004.

[13] R. Ospina, F. Cribari-Neto e K. L. P. Vasconcellos: *Improved point and intervalar estimation for a beta regression model*. Computational Statistics & Data Analysis, 51:960–981, 2006.

[14] R. Ospina e S. L. P. Ferrari: *A general class of zero-or-one inflated beta regression models*. Computational Statistics & Data Analysis, 56:1609–1623, 2012.

[15] W. Press, S. Teukolsky, W. Vetterling e B. Flannery: *Numerical recipes in C: The art of scientific computing*. Cambridge University Press, 1992.

[16] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[17] J. A. Rogers, D. Polhamus, W. R. Gillespie, K. Ito, K. Romero, Q. R., D. Stephenson, M. R. Gastonguay e B. Corrigan: *Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis*. Journal of Pharmacokinetics and Pharmacodynamics, 39(5):479–498, 2012.

[18] A. B. Simas, W. Barreto-Souza e R. A. V.: *Improved estimators for a general class of beta regression models*. Computational Statistics & Data Analysis, 2:348–366, 2010.