EFEITOS DA ESPECIFICAÇÃO INCORRETA DAS FUNÇÕES DE LIGAÇÃO NO MODELO DE REGRESSÃO BETA COM DISPERSÃO VARIÁVEL

Diego Ramos CANTERLE[1]
Bruna Gregory PALM[2]
Fábio Mariano BAYER[3]

▪ RESUMO: O modelo de regressão beta com dispersão variável é utilizado para modelar dados contínuos no intervalo (0,1), assumindo distribuição beta para a variável de interesse. Nesse modelo, são consideradas estruturas de regressão para os parâmetros de média e de dispersão, que envolvem covariáveis, parâmetros desconhecidos e funções de ligação. Este trabalho aborda o problema da má especificação da função de ligação no submodelo da dispersão do modelo de regressão beta com dispersão variável. Por meio de simulações de Monte Carlo, considerando diferentes funções de ligação, foram avaliadas as taxas de cobertura e balanceamento dos parâmetros do submodelo da média, assim como dos valores preditos da média e da dispersão. Também foi avaliado o viés relativo percentual dos estimadores desses modelos. Verificou-se que a especificação incorreta da função de ligação do submodelo da dispersão influencia consideravelmente nas inferências do modelo. Por fim, realizou-se um estudo com dados reais.

▪ PALAVRAS-CHAVE: Função de ligação; intervalos de confiança; modelo de regressão beta com dispersão variável; simulações de Monte Carlo

1 Introdução

O modelo de regressão beta, introduzido por Ferrari e Cribari-Neto (2004), é de ampla utilização prática quando necessita-se modelar variáveis pertencentes

[1]Universidade Federal de Santa Maria, Bacharelado de Estatística, CEP: 97105-900, Santa Maria, RS, Brasil. E-mail: *diegocanterle@gmail.com*
[2]Universidade Federal de Santa Maria, Mestrado em Engenharia de Produção, CEP: 97105-900, Santa Maria, RS, Brasil. E-mail: *brunagpalm@gmail.com*
[3]Universidade Federal de Santa Maria, Departamento de Estatística e LACESM, CEP: 97105-900, Santa Maria, RS, Brasil. E-mail: *bayer@ufsm.br*

ao intervalo contínuo (0,1). Nestes modelos, assume-se que a variável dependente Y possui distribuição beta, e que a média de Y é modelada por meio de uma estrutura de regressão que envolve parâmetros desconhecidos, covariáveis e uma função de ligação. Uma extensão do modelo seminal proposto por Ferrari e Cribari-Neto (2004) é o modelo de regressão beta com dispersão variável, já discutido por Smithson e Verkuilen (2006), Simas et al. (2010), Ferrari e Pinheiro (2011) e Bayer e Cribari-Neto (2015). Neste modelo, o parâmetro de dispersão (ou precisão) de Y é modelado por meio de uma estrutura de regressão da mesma forma que a média. Recentes aplicações a dados reais do modelo de regressão beta com dispersão variável podem ser vistos em Silva e Souza (2014), Cribari-Neto e Souza (2013) e Souza e Cribari-Neto (2013).

Na seleção de modelos de regressão beta duas importantes decisões são tipicamente necessárias: (i) seleção das covariáveis regressoras e (ii) escolha das funções de ligação. A seleção das covariáveis importantes nas estruturas de regressão da média e da dispersão é um tópico amplamente explorado em Bayer e Cribari-Neto (2015). A correta especificação das funções de ligação também é algo que merece atenção, mas é usualmente negligenciado em estudos práticos. A correta modelagem do parâmetro de dispersão, por meio da seleção adequada de covariáveis e da função de ligação, tem implicações na eficiência dos estimadores dos parâmetros da estrutura de regressão da média (Smyth e Verbyla, 1999; Bayer e Cribari-Neto, 2015). Além de melhorar as inferências sobre os parâmetros da estrutura da média, em diversas aplicações, há o interesse direto na modelagem da dispersão com objetivo de identificar fontes da variabilidade das observações (Smyth e Verbyla, 1999; Wu e Li, 2012).

Como forma de quantificar o impacto da especificação incorreta da função de ligação da estrutura da média de Y sobre as inferências no modelo de regressão beta com dispersão constante, Andrade (2007) realiza um extenso estudo de simulação. Para testar a correta especificação da função de ligação no modelo de regressão beta, Oliveira (2013) avalia o desempenho do teste RESET. De forma semelhante, Pereira e Cribari-Neto (2013) avaliam o teste RESET no modelo de regressão beta inflacionado. Ramalho et al. (2011) propõe a utilização do teste RESET e outras alternativas para a seleção da função de ligação no modelo de regressão beta. Contudo, nenhum desses trabalhos avalia o impacto do erro de especificação da função de ligação da estrutura da dispersão no modelo de regressão beta com dispersão variável.

Neste sentido, o presente trabalho tem o objetivo de avaliar o impacto da especificação incorreta da função de ligação do submodelo da dispersão no modelo de regressão beta com dispersão variável. Serão consideradas simulações de Monte Carlo avaliando dois aspectos: (i) taxa de cobertura e balanceamento dos parâmetros da estrutura do submodelo da média (β), assim como para a média (μ) e a dispersão de Y (σ) e (ii) viés relativo dos estimadores de β, μ e σ. Para cumprir estas etapas a variável resposta com distribuição beta será gerada assumindo funções de ligação (para média e dispersão) conhecidas e o modelo será então ajustado usando as funções de ligação corretas e algumas funções de ligação incorretas para

a estrutura da dispersão.

Este trabalho está organizado da seguinte forma. A Seção 2 apresenta o modelo de regressão beta com dispersão variável, juntamente com uma breve descrição das funções de ligação consideradas. A Seção 3 apresenta os resultados numéricos e discussões. Na Seção 4 é apresentado um estudo em dados reais. Finalmente, na Seção 5 são apresentadas as conclusões do trabalho.

2 O modelo de regressão beta com dispersão variável

A densidade beta é definida 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 $\Gamma(u) = \int_0^\infty t^{u-1}e^{-t}\mathrm{dt}$. Desta forma, uma variável aleatória Y que possui distribuição beta apresenta média e variância 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 em Ferrari e Cribari-Neto (2004) consiste em uma reparametrização da densidade beta dada em (1). A parametrização proposta é indexada pelos parâmetros de média (μ) e de precisão (ϕ), de tal forma que: $\mu = p/(p+q)$ e $\phi = p+q$, portanto $p = \mu\phi$ e $q = (1-\mu)\phi$. Sendo assim:

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

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

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

onde $0 < \mu < 1$ e $\phi > 0$.

O modelo de regressão beta proposto em Ferrari e Cribari-Neto (2004) considera um parâmetro de precisão, ϕ, constante ao longo das observações. Contudo, ao assumirmos ϕ constante erroneamente as perdas de eficiência dos estimadores podem ser substanciais, como discutido em Bayer e Cribari-Neto (2015). No modelo de regressão beta com dispersão variável, o parâmetro de precisão é assumido variável ao longo das observações, sendo modelado em termos de covariáveis, parâmetros desconhecidos e uma função de ligação, da mesma forma que a média.

Neste trabalho, assim como em Cribari-Neto e Souza (2012) e Bayer e Cribari-Neto (2015), é considerada uma outra reparametrização para a densidade beta. Não é considerado o parâmetro de precisão ϕ, mas sim um parâmetro de dispersão σ, dado por $\sigma^2 = \frac{1}{1+\phi}$; ou seja, $\phi = \frac{1-\sigma^2}{\sigma^2}$. Desta maneira, a densidade beta pode ser escrita da seguinte forma:

$$f(y;\mu,\sigma) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\mu\frac{1-\sigma^2}{\sigma^2}\right)\Gamma\left((1-\mu)\frac{1-\sigma^2}{\sigma^2}\right)} y^{\mu\frac{1-\sigma^2}{\sigma^2}-1}(1-y)^{(1-\mu)\frac{1-\sigma^2}{\sigma^2}-1}, \qquad (3)$$

em que $0 < \mu < 1$ e $0 < \sigma < 1$. Nesta parametrização os dois parâmetros que indexam a densidade assumem valores no intervalo unitário padrão (0,1), tendo assim as mesmas opções de funções de ligação para as duas estruturas de regressão. A parametrização utilizando σ ao invés de ϕ tem a vantagem de ter o espaço paramétrico limitado superiormente e inferiormente, tornando mais fácil a identificação de grandes (próxima a um) e pequenas (próxima de zero) variabilidades.

Sendo $Y_1, \ldots, Y_n$ variáveis aleatórias independentes, em que cada Y_t, $t = 1, \ldots, n$, tem distribuição dada por (3), com média μ_t e dispersão σ_t, define-se o modelo de regressão beta com dispersão variável por:

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

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 dispersão, respectivamente, $r + s = k < n$, $\eta_{1t} = x_t^\top\beta$ e $\eta_{2t} = z_t^\top\gamma$ são os preditores lineares, $x_t^\top = (x_{t1}, \ldots, x_{tr})$ e $z_t^\top = (z_{t1}, \ldots, z_{ts})$ representam as t-ésimas observações das 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 $g_i(0,1) \to \mathbb{R}$, para $i = 1,2$ (Cribari-Neto e Souza, 2012; Bayer e Cribari-Neto, 2015).

Para a estimação dos parâmetros 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:

$$\ell(\beta,\gamma) = \sum_{t=1}^{n} \ell_t(\mu_t,\sigma_t), \qquad (4)$$

em que

$$\ell_t(\mu_t,\sigma_t) = \log\Gamma\left(\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \log\Gamma\left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2}\right) - \log\Gamma\left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2}\right)$$
$$+ \left(\mu_t\frac{1-\sigma_t^2}{\sigma_t^2} - 1\right)\log y_t + \left((1-\mu_t)\frac{1-\sigma_t^2}{\sigma_t^2} - 1\right)\log(1-y_t).$$

Para a maximização da função dada em (4) se faz necessário o uso de algoritmos de otimização não-linear. Usualmente, utiliza-se o método quasi-Newton BFGS (Press et al., 1992). Para mais detalhes inferenciais e expressões matriciais do vetor escore e matriz informação de Fisher ($I_F(\beta,\gamma)$), ver Cribari-Neto e Souza (2012).

Sob condições usuais de regularidade e em grandes amostras 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.

O intervalo de confiança para os parâmetros dos modelos, sendo $\theta = (\beta^\top,\gamma^\top)^\top$, é definido por:

$$[\widehat{\theta}_i - \Phi^{-1}(1-\alpha/2)\widehat{\text{ep}}(\widehat{\theta}_i); \widehat{\theta}_i + \Phi^{-1}(1-\alpha/2)\widehat{\text{ep}}(\widehat{\theta}_i)],$$

em que os erros padrão de $\widehat{\theta}$ são dados por $\widehat{\text{ep}}(\widehat{\theta}) = [\text{diag}(\widehat{\text{cov}}(\widehat{\theta}))]^{1/2}$, sendo $\widehat{\text{cov}}(\widehat{\theta}) = I_F(\widehat{\beta},\widehat{\gamma})^{-1}$ e Φ é a função distribuição acumulada da distribuição normal padrão de forma que Φ^{-1} é a função quantil.

Os intervalos de confiança para μ e σ, i = 1,2 respectivamente, são dados por:

$$[g_i^{-1}(\widehat{\eta}_{it} - \Phi^{-1}(1-\alpha/2)\widehat{\text{ep}}(\widehat{\eta_{it}})); g_i^{-1}(\widehat{\eta}_{it} + \Phi^{-1}(1-\alpha/2)\widehat{\text{ep}}(\widehat{\eta_{it}}))],$$

em que os erros padrão de $\widehat{\eta_{it}}$, para i = 1,2, são estimados por $\widehat{\text{ep}}(\widehat{\eta_{1t}}) = (x_t\widehat{\text{cov}}(\widehat{\beta})x_t^\top)^{1/2}$ e $\widehat{\text{ep}}(\widehat{\eta_{2t}}) = (z_t\widehat{\text{cov}}(\widehat{\gamma})z_t^\top)^{1/2}$.

2.1 Funções de ligação

No modelo de regressão beta com dispersão variável diferentes funções de ligação podem ser consideradas, tanto na estrutura da média quanto na da dispersão. Quando é considerada a parametrização utilizando σ, as mesmas opções para funções de ligação podem ser usadas em ambos submodelos. As funções de ligação possíveis são aquelas que associam os valores de μ_t ou σ_t, do intervalo padrão (0,1), aos valores dos preditores lineares η_{1t} e η_{2t}, na reta real. As funções de ligação usuais nesses casos são logit, probit, loglog e Cauchy (Koenker e Yoon, 2009), dadas, respectivamente, por:

$$\begin{aligned} g(x) &= \log\left(\frac{x}{1-x}\right), \\ g(x) &= \Phi^{-1}(x), \\ g(x) &= -\log[-\log(x)], \\ g(x) &= \tan[\pi(x-0{,}5)]. \end{aligned}$$

A Figura 1 apresenta uma comparação gráfica das funções de ligação consideradas nesse estudo. Conforme discutido em Canterle e Bayer (2015), pode-se

observar que a função de ligação loglog é assimétrica; a função de ligação Cauchy tem as caudas mais pesadas e a função de ligação probit possui as caudas mais leves. Com $\eta \approx 0$, as funções de ligação logit, probit e Cauchy levam a valores muito próximos de μ. Também pode-se verificar que todas as funções são praticamente indistinguíveis quando valores de μ são muito próximos dos extremos do intervalo padrão (0,1).

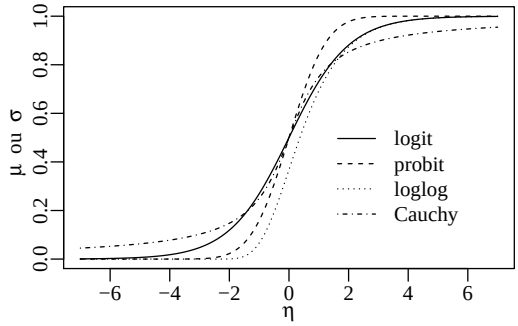


Figura 1 - Gráfico comparativo entre as funções de ligação.

3 Resultados numéricos

Esta seção apresenta resultados numéricos das simulações de Monte Carlo utilizadas na avaliação dos efeitos da especificação incorreta da função de ligação sobre as inferências do modelo. São considerados as taxas de cobertura (TC) e o balanceamento (B) dos intervalos de confiança, com $\alpha = 0{,}05$, dos parâmetros do submodelo da média (β), dos parâmetros da média μ_t e da dispersão σ_t. Além disso, foram analisados os vieses relativos médios para β, μ e σ, a fim de avaliar a influência da falta de especificação da função de ligação no viés das estimativas. Foram consideradas $R = 50.000$ réplicas de Monte Carlo e o tamanho amostral considerado foi de $n = 500$. Em cada réplica de Monte Carlo gerou-se n ocorrências da variável aleatória Y_t com função densidade dada em (3), parâmetro de média definida por $\mu_t = g_1^{-1}(\eta_{1t})$ e parâmetro de dispersão definido por $\sigma_t = g_2^{-1}(\eta_{2t})$, em que:

$$\eta_{1t} = \beta_0 + \beta_1 x_{1t},$$
$$\eta_{2t} = \gamma_0 + \gamma_1 z_{1t} + \gamma_2 z_{2t}.$$

As covariáveis x_1, z_1 e z_2 foram geradas da distribuição uniforme (0,1) e constantes durante todas as réplicas. Para todos os modelos foram utilizados $\beta_0 = \beta_1 = -2$. Na geração dos dados, para $g_1(\cdot)$ foi considerada função de ligação logit e para $g_2(\cdot)$ as funções de ligação logit, probit, loglog e Cauchy. Foram considerados cenários para dispersão com (i) $\sigma \in [0{,}01;0{,}24]$ e com (ii) σ

$\in [0{,}02;0{,}47]$. Para o cenário (i) temos: $\gamma = (-3; 2; -2)^\top$, $\gamma = (-1{,}6; 0{,}95; -0{,}75)^\top$, $\gamma = (-0{,}7; 0{,}4; -0{,}7)^\top$ e $\gamma = (-2; 1{,}2; -14)$ quando os dados são gerados com as funções de ligação logit, probit, loglog e Cauchy, respectivamente, para o submodelo da dispersão. Já para o cenário (ii) temos: $\gamma = (-2; 2; -2)^\top$, $\gamma = (-0{,}8; 0{,}8; -1{,}2)^\top$, $\gamma = (-0{,}7; 1; -0{,}7)^\top$ e $\gamma = (-1{,}7; 1{,}9; -14)$ para as funções de ligação logit, probit, loglog e Cauchy, respectivamente. Todas as implementações computacionais foram desenvolvidas em linguagem `R` (R Development Core Team, 2014). Para os ajustes dos modelos foi utilizada a função `gamlss` (Rigby e Stasinopoulos, 2005). Cabe destacar que outros cenários e tamanhos amostrais foram considerados, mas por questões de brevidade e de semelhança de resultados foram suprimidos.

A TC é definida pela proporção de réplicas de Monte Carlo do intervalo de confiança que conteve o parâmetro dentro do número total de réplicas considerado (Oliveira e Ferrari, 2004; Lemonte e Silva, 2006). O B é dado pela subtração entre a proporção de vezes em que o valor do parâmetro é maior que o limite superior do intervalo de confiança e a proporção de vezes em que é menor que o limite inferior do intervalo (Andrade, 2007). Para estudo do viés nas estimativas das médias μ_t, das dispersões σ_t e dos parâmetros β_i consideramos o viés relativo (VR) médio, dado da seguinte forma:

$$\mathrm{VR}_\mu = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{\mu_t} - \mu_t|}{\mu_t},$$

$$\mathrm{VR}_\sigma = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{\sigma_t} - \sigma_t|}{\sigma_t},$$

$$\mathrm{VR}_\beta = \frac{1}{r}\sum_{i=1}^{r}\frac{|\hat{\beta_i} - \beta_i|}{|\beta_i|},$$

em que $\hat{\mu_t}$ e $\hat{\sigma_t}$ são as estimativas para a média e para a dispersão, respectivamente, para as n observações e $\hat{\beta_i}$ são as estimativas dos r β_i. Para TC, espera-se que o valor aproxime-se de $(1-\alpha)$, ou seja 0,95, enquanto que para B e VR espera-se que sejam próximos de zero. Para μ e σ temos n valores de TC, sendo uma para cada μ_t e σ_t. Neste sentido, serão apresentados a média, o máximo e o mínimo dos n valores de TC dos μ_t e σ_t.

Na Tabela 1 encontram-se os resultados numéricos da avaliação da falta de especificação da função de ligação de $g_2(\cdot)$ sobre as inferências de μ_t. Em negrito estão destacados os melhores resultados. Nota-se a melhora nos valores de TC e B quando as funções de ligação são iguais na geração e na estimação do modelo. Quando as funções de ligação são as mesmas na geração e na estimação as TC são próximas de 95% e as medidas de balanceamento de zero. A escolha da função de ligação Cauchy tanto na estimação quando na geração da amostra mostrou valores maiores em todas as medidas consideradas nos distintos cenários. As maiores distorções nos valores de TC foram quando assumiu-se erroneamente as funções de ligação probit e loglog em vez da Cauchy, como pode ser verificado no segundo cenário de σ. Os valores de TC mínimo chegaram a 89,57% e 89,08% para as

funções de ligação probit e loglog, respectivamente. A maior distorção de B pode-se verificar no mesmo cenário, o valor máximo de B chegou a 1,13% para a probit e 1,29% para a loglog.

Tabela 1 - Taxa de cobertura e balanceamento percentual de μ_t

	$\sigma \in$ [0,01;0,24]						$\sigma \in$ [0,02;0,47]					
	Taxa de cobertura e balanceamento percentual de μ_t											
	TC			B			TC			B		
	média	max	min	média	max	min	média	max	min	média	max	min
	gerado logit											
logit	**94,63**	**94,76**	**94,46**	0,08	0,23	**-0,08**	**94,66**	**94,90**	**94,45**	**-0,02**	**0,29**	**-0,26**
probit	93,07	93,48	92,57	0,12	0,30	-0,14	93,20	93,40	92,96	0,11	0,40	-0,30
loglog	91,27	91,94	90,42	0,22	0,48	-0,13	90,90	91,62	90,25	0,53	0,99	-0,62
Cauchy	97,94	98,22	97,64	-0,05	0,14	-0,27	97,32	97,70	96,98	-0,84	-0,59	-0,99
	gerado probit											
logit	95,48	95,68	95,36	0,04	0,13	-0,03	95,47	95,67	95,29	-0,06	0,18	-0,31
probit	**94,69**	**94,85**	94,49	0,12	0,23	-0,09	**94,73**	**94,79**	94,60	**0,03**	0,28	-0,30
loglog	93,63	93,94	93,26	0,21	0,39	-0,18	93,33	93,70	92,82	0,32	0,68	-0,21
Cauchy	97,32	97,63	97,04	-0,01	0,3	-0,03	96,76	97,13	96,44	-1,03	-0,59	-1,26
	gerado loglog											
logit	95,43	95,66	95,22	0,06	0,23	-0,27	96,40	96,66	95,85	-0,27	0,04	-0,52
probit	95,09	95,31	94,88	0,10	0,29	-0,25	95,85	96,00	95,76	-0,14	0,11	-0,36
loglog	**94,66**	**94,86**	94,50	0,13	0,36	**-0,18**	**94,60**	**94,75**	**94,36**	**0,02**	0,31	**-0,25**
Cauchy	96,42	96,81	96,09	0,00	0,38	-0,73	96,85	97,36	96,44	-1,48	-0,83	-1,76
	gerado Cauchy											
logit	92,90	93,30	92,34	0,09	0,66	-0,38	91,18	92,03	90,13	0,04	0,96	-0,73
probit	92,74	93,21	92,10	0,08	0,69	-0,44	90,80	91,79	89,57	0,05	1,13	-0,82
loglog	92,61	93,14	91,89	0,11	0,77	-0,46	90,52	91,59	89,08	0,18	1,29	-0,74
Cauchy	**94,81**	**94,93**	**94,71**	**0,07**	**0,32**	**-0,12**	**94,79**	**94,93**	**94,64**	**0,02**	**0,18**	**-0,18**

A Tabela 2 apresenta os resultados numéricos da avaliação do efeito da incorreta especificação da função de ligação $g_2(\cdot)$ nas inferências sobre σ_t. Nota-se o grande impacto da especificação incorreta. No caso de incorreta especificação da função de ligação, verificam-se taxas de cobertura médias muito menores que o valor nominal esperado e taxas de cobertura mínimas que chegam a zero. Por exemplo, no primeiro cenário de σ verifica-se que quando os dados são gerados com a função de ligação Cauchy e o modelo é estimado com a logit, a TC média é de 34,44% e a mínima é de 0,00%, valores muito abaixo do esperado. Também verifica-se valor para B de 100,00%, quando espera-se que fique perto de zero.

Analisando a Tabela 3, com resultados do efeito da incorreta especificação de $g_2(\cdot)$ sobre inferências de β, observa-se que o erro na função de ligação causa impacto, também, nas inferências do submodelo da média. Observa-se que na maior parte dos casos, em exceção aos dados gerados pela função de ligação loglog com $\sigma \in [0{,}01; 0{,}24]$, os valores de taxa de cobertura dos parâmetros do modelo da média sofrem impacto pela incorreta especificação da função de ligação da estrutura da dispersão. Por exemplo, para dados gerados pela Cauchy e estimados pela loglog apresentam TC igual a 89,74% e B de 1,29%, enquanto que quando estimados pela

Cauchy os valores de TC e B são iguais a 94,82% e −0,17%, respectivamente.

Tabela 2 - Taxa de cobertura e balanceamento percentual de σ_t

	$\sigma \in$ [0,01;0,24]						$\sigma \in$ [0,02;0,47]					
	TC			B			TC			B		
	média	max	min	média	max	min	média	max	min	média	max	min
	gerado logit											
logit	**94,79**	**95,07**	**94,54**	1,55	**2,55**	**0,24**	**94,74**	**95,19**	**94,36**	1,70	**2,73**	**0,28**
probit	84,66	94,99	20,20	2,07	79,80	-29,53	85,73	94,85	36,32	1,60	63,67	-28,03
loglog	64,30	94,75	0,00	-0,16	100,00	-83,19	60,25	94,69	0,00	-3,84	100,00	-89,25
Cauchy	21,61	96,61	0,00	-1,21	100,00	-100,00	29,75	95,52	0,00	-4,30	99,96	-100,00
	gerado probit											
logit	88,27	95,35	52,93	2,31	26,59	-47,07	89,07	95,32	48,05	2,22	22,90	51,94
probit	**94,72**	**95,01**	**94,32**	1,75	**3,07**	**0,33**	**94,73**	95,15	**94,25**	**1,76**	**3,00**	**0,33**
loglog	88,74	94,99	36,56	2,46	63,44	-17,83	84,31	95,12	15,01	2,68	84,99	-29,35
Cauchy	25,93	96,51	0,00	1,32	100,00	-100,00	35,47	95,53	0,00	-3,27	99,56	-100,00
	gerado loglog											
logit	83,10	95,55	17,78	2,40	41,58	-82,22	70,07	95,60	0,15	1,77	69,39	-99,85
probit	91,44	95,36	76,85	2,55	17,55	-22,97	84,09	95,45	34,73	2,33	37,71	65,27
loglog	**94,70**	**95,09**	**94,05**	**1,89**	**3,22**	**0,69**	**94,68**	**95,09**	**93,98**	1,78	**3,49**	**0,17**
Cauchy	37,40	96,13	0,00	2,81	99,88	-100,00	33,87	95,48	0,00	-2,38	99,71	-100,00
	gerado Cauchy											
logit	34,44	93,95	0,00	-21,45	100,00	-100,00	24,44	93,36	0,00	-29,82	99,97	-99,97
probit	30,23	93,65	0,00	-24,38	100,00	-100,00	21,55	92,79	0,00	-33,79	99,97	-99,97
loglog	27,01	92,88	0,00	-26,32	100,00	-100,00	19,24	92,20	0,00	-36,44	99,97	-99,97
Cauchy	**95,01**	**95,29**	**94,61**	**0,82**	**1,98**	**-0,14**	**95,01**	**95,36**	**94,68**	**0,79**	**1,96**	**-0,52**

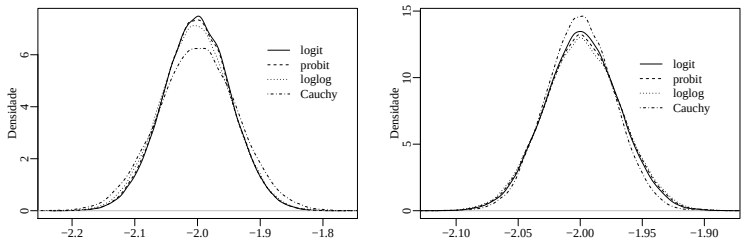
Essa distorção dos intervalos de confiança para β está diretamente relacionada com a perda de eficiência dos estimadores sob incorreta especificação do submodelo de σ_t. A Figura 2 evidencia esse fato, mostrando as densidades empíricas estimadas de $\widehat{\beta_1}$ quando consideradas diferentes funções de ligação para a dispersão $g_2(\cdot)$. Nota-se, por exemplo, na Figura 2(a), que quando a função de ligação correta da estrutura de σ_t é a logit, mas estima-se o modelo incorretamente com a Cauchy, a distribuição de $\widehat{\beta_1}$ apresenta maior variabilidade.

A Tabela 4 apresenta os viéses relativos percentuais médios dos estimadores de β, μ e σ. Pode-se observar que o erro da função de ligação do submodelo da dispersão acarreta em viés desprezível nos estimadores de $\hat{\beta}$ e de $\hat{\mu}$. Por outro lado, errar essa função de ligação implica em alto viés dos $\hat{\sigma}_t$, chegando a 29,95% de viés quando gerado da logit e estimada com a Cauchy. De forma geral, quando o modelo é estimado erroneamente com a função de ligação Cauchy temos os maiores viéses. Ainda, quando os dados são originários da função de ligação Cauchy, estimar o modelo com as outras funções de ligação estudadas causam os maiores viéses. Os altos viéses de $\hat{\sigma}_t$ sob incorreta especificação da função de ligação podem explicar grande parte dos resultados distorcidos de TC sobre σ_t, evidenciados na Tabela 2.

De forma geral, pode-se verificar que o balanceamento apresenta maiores distorções nos casos em que os dados são gerados pela função de ligação Cauchy

Tabela 3 - Taxa de cobertura e balanceamento percentual dos β's.

	$\sigma \in$ [0,01;0,24]				$\sigma \in$ [0,02;0,47]			
	$\beta_0 = -2$		$\beta_1 = -2$		$\beta_0 = -2$		$\beta_1 = -2$	
	TC	B	TC	B	TC	B	TC	B
	gerado logit							
logit	**94,54**	**0,05**	**94,63**	**0,17**	**94,45**	**-0,23**	**94,53**	**0,13**
probit	92,96	-0,08	93,35	0,20	92,98	-0,30	93,32	0,27
loglog	91,01	-0,08	91,81	0,38	90,69	-0,63	91,40	0,89
Cauchy	97,65	-0,23	97,58	-0,25	97,08	-0,75	96,86	-0,10
	gerado probit							
logit	95,40	0,00	95,43	0,07	95,49	-0,29	95,40	0,04
probit	94,58	0,08	**94,74**	0,17	**94,73**	-0,26	**94,80**	**0,00**
loglog	93,62	-0,18	93,94	0,32	93,23	-0,02	93,82	0,27
Cauchy	97,04	-0,36	96,95	0,40	96,65	-1,05	96,38	0,05
	gerado loglog							
logit	95,34	-0,23	95,19	0,12	96,23	-0,01	96,13	-0,38
probit	95,01	-0,17	94,95	0,15	95,79	0,00	95,74	-0,25
loglog	94,54	**-0,16**	94,62	**0,10**	**94,66**	-0,19	**94,72**	**0,04**
Cauchy	96,21	-0,73	95,93	0,54	96,80	-0,84	96,42	-0,76
	gerado Cauchy							
logit	92,61	0,48	93,12	-0,41	90,60	0,96	91,74	-1,06
probit	92,42	0,53	92,95	-0,44	90,14	1,14	91,44	-1,10
loglog	92,18	0,61	92,92	-0,49	89,74	1,29	91,20	-1,17
Cauchy	**94,85**	**-0,07**	**94,83**	**0,04**	**94,82**	**-0,17**	**95,01**	**-0,01**



(a) gerado logit (b) gerado Cauchy

Figura 2 - Gráficos das densidades de $\widehat{\beta_1}$ considerando diferentes funções de ligação.

e considera-se as outras funções de ligação para estimação. Isso pode ocorrer pelo fato da função de ligação Cauchy possuir caudas mais pesadas, e maior diferença entre as funções de ligação nos intervalos de σ considerados. Ainda, as maiores distorções das taxas de cobertura ocorrem quando os dados são originários das funções de ligação Cauchy e logit, com maiores dispersões. Assim como para B e TC, os maiores valores de viés relativo para σ são verificados quando os dados são gerados pela Cauchy. Para μ e β, os viéses absolutos médios são despresíveis, sendo inferiores a 0,1%.

Tabela 4 - Viés relativo percentual.

	$\sigma \in$ [0,01;0,24]			$\sigma \in$ [0,02;0,47]		
	β	μ	σ	β	μ	σ
	gerado logit					
logit	**0,00**	**0,00**	**0,36**	**0,01**	**0,02**	**0,00**
probit	0,01	0,01	4,01	0,03	0,03	3,55
loglog	0,01	0,01	7,67	0,08	0,08	7,77
Cauchy	0,04	0,02	29,95	0,03	0,28	22,14
	gerado probit					
logit	0,00	0,00	3,38	0,00	0,03	2,94
probit	0,01	**0,00**	**0,36**	0,01	**0,02**	**0,36**
loglog	0,01	0,01	2,93	0,04	0,04	3,56
Cauchy	0,06	0,03	25,28	0,11	0,43	19,49
	gerado loglog					
logit	0,01	0,01	4,49	0,06	0,09	6,74
probit	0,01	0,01	2,36	0,03	0,05	3,93
loglog	**0,01**	**0,01**	**0,37**	**0,02**	**0,01**	**0,36**
Cauchy	0,11	0,06	17,71	0,23	0,64	23,85
	gerado Cauchy					
logit	0,03	0,02	15,02	0,05	0,04	22,35
probit	0,04	0,03	17,10	0,06	0,04	25,77
loglog	0,04	0,03	18,91	0,06	0,04	28,85
Cauchy	**0,00**	**0,00**	**0,35**	**0,00**	**0,00**	**0,35**

Assim, verifica-se que a má especificação da função de ligação no submodelo da dispersão acarreta em perda de eficiência dos estimadores da média, dos parâmetros do modelo da média e da dispersão, assim como viés nos $\hat{\sigma}_t$. Os principais problemas encontrados estão nos estimadores de σ, portanto, quando também interessa-se pelo comportamento da dispersão, como salientado por Smyth e Verbyla (1999), deve-se ter uma atenção especial na seleção da função de ligação. Para verificação da correta especificação das funções de ligação recomenda-se o uso de testes RESET (Ramsey, 1969), conforme sugerido por Canterle e Bayer (2015) em modelos lineares generalizados para dados binários, por Oliveira (2013) no modelo de regressão beta com dispersão constante e por Pereira e Cribari-Neto (2013) no modelo de regressão beta inflacionada.

4 Análise em dados reais

Esta seção apresenta uma aplicação a dados reais com intuito de avaliar empiricamente os efeitos da falta de especificação da função de ligação nos modelos de regressão beta com dispersão variável. Os dados considerados são referentes a um estudo sobre crença religiosa e inteligência em 124 países apresentado em Cribari-Neto e Souza (2013). A variável dependente, y, é a proporção de descrentes em cada país. As covariáveis consideradas são: a média do coeficiente de inteligência da população de cada país (IQ), IQ ao quadrado (IQ^2), uma variável *dummy* onde

recebe 1 se o percentual de muçulmanos é maior do que 50% e 0 caso contrário ($MUSL$), a renda per capita ajustada pela paridade do poder de compra em 2008, em mil dólares ($INCOME$), o quadrado do quociente entre a soma das importações e exportações e o Produto Interno Bruto em 2008 ($OPEN^2$) e o logaritmo do quociente entre a soma das importações e exportações e o Produto Interno Bruto em 2008 ($logOPEN$).

A seleção da função de ligação é realizada por meio de testes RESET considerando a estatística da razão de verossimilhanças (Neyman e Pearson, 1928). Para seleção das covariáveis nas estruturas de regressão da média e da dispersão, foi considerado um extensivo trabalho de ajustes e análises de diagnóstico. De acordo com o teste RESET aplicado aos dados de proporção de descrentes, utilizar a função de ligação logit tanto para o submodelo da média como para o da dispersão acarreta em incorreta especificação do modelo ($p-$valor = 0,0010). Quando considerada a função de ligação loglog para o submodelo da média e a logit para o submodelo da dispersão o teste RESET sugere correta especificação do modelo ($p-$valor = 0,8016). Os modelos corretamente especificado e incorretamente especificado, segundo o teste RESET, estão apresentados nas Tabelas 5 e 6, respectivamente.

Tabela 5 - Parâmetros estimados e p-valores com ligação loglog para a média e logit para dispersão

	$\hat{\beta}$	$p-$valor	$\hat{\gamma}$	$p-$valor
Intercepto	5,8650	$< 1 \times 10^{-4}$	$-9{,}6759$	$< 1 \times 10^{-4}$
IQ	$-0{,}2070$	$< 1 \times 10^{-4}$	0.0610	$< 1 \times 10^{-4}$
IQ^2	0,0014	$< 1 \times 10^{-4}$	-	-
$MUSL$	$-0{,}1407$	$< 1 \times 10^{-4}$	-0.8751	$< 1 \times 10^{-4}$
$INCOME$	0,0068	0,0149	-	-
$OPEN^2$	$9{,}067 \times 10^{-6}$	0,0398	-	-
$logOPEN$	-	-	0,7230	$< 1 \times 10^{-4}$

Tabela 6 - Parâmetros estimados e p-valores com ligação logit para a média e logit para dispersão

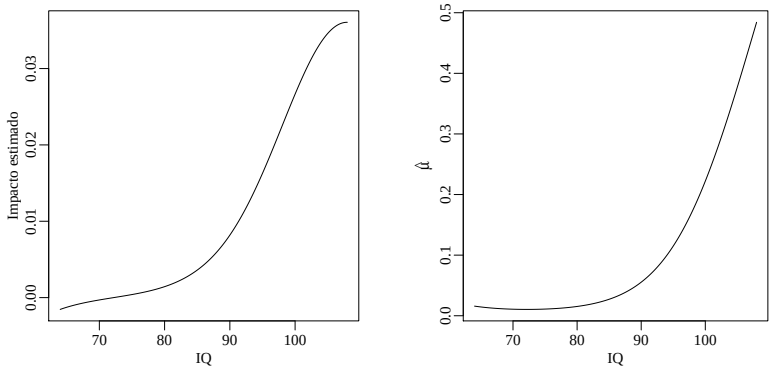
	$\hat{\beta}$	$p-$valor	$\hat{\gamma}$	$p-$valor
Intercepto	9,3260	0,0052	$-9{,}3076$	$< 1 \times 10^{-4}$
IQ	$-0{,}4129$	$< 1 \times 10^{-4}$	0.0614	$< 1 \times 10^{-4}$
IQ^2	0,0036	$< 1 \times 10^{-4}$	-	-
$MUSL$	$-0{,}1725$	$< 1 \times 10^{-4}$	-1.0169	$< 1 \times 10^{-4}$
$INCOME$	0,0146	0,0076	-	-
$OPEN^2$	$1{,}929 \times 10^{-5}$	0,0805	-	-
$logOPEN$	-	-	0,6406	$< 1 \times 10^{-4}$

Analisando a Tabela 5, pode-se verificar que todos os parâmetros foram significativos ao nível de 5%. Comparando os resultados obtidos em Cribari-Neto e Souza (2013) com os da Tabela 5, verifica-se que os modelos da média são semelhantes, porém no presente trabalho é considerado $OPEN^2$ no lugar de $logOPEN$. A função de ligação na estrutura de regressão da média é a mesma (loglog) e as variáveis IQ e $MUSL$ influenciam negativamente na proporção de descrentes em ambos os casos.

Pode-se verificar que as variáveis $INCOME$ e $OPEN^2$ influenciam positivamente na média da variável Y, enquanto que $MUSL$ influencia negativamente. Tais influencias não são facilmente interpretadas para a variável índice de inteligência, pois IQ apresenta coeficiente negativo, enquanto que IQ^2 apresenta coeficiente positivo. Desta forma, torna-se necessário estimar o impacto da inteligência na proporção média de descrentes, como realizado por Cribari-Neto e Souza (2013). Esse impacto é definido como:

$$\frac{\partial E(y_t)}{\partial IQ_t} = \frac{\partial g^{-1}(\eta_t)}{\partial IQ_t} = \frac{\partial \mu_t}{\partial IQ_t},$$

sendo $\mu_t = g^{-1}(\beta_0 + \beta_1 IQ + \beta_2 IQ^2 + \beta_3 MUSL + \beta_4 INCOME + \beta_5 OPEN^2)$. Para as variáveis $INCOME$ e $OPEN^2$ foram considerados seus valores medianos, para $MUSL$ foi considerado zero, ou seja, países não Muçulmanos. A variável IQ recebeu valores de 64 a 108, com o intuito de verificar o impacto de sua variação em relação a proporção de descrentes. A Figura 3(a) apresenta o impacto estimado da inteligência média na proporção de descrentes e a Figura 3(b) mostra a relação da proporção de descrentes com o IQ, fixando as demais variáveis como anteriormente. Percebe-se o impacto de IQ na proporção de descrentes é sempre positivo. Verifica-se também que a média de IQ aumenta com o aumento da proporção no número de descrentes.



(a) IQ versus impacto estimado. (b) IQ versus $\hat{\mu}$.

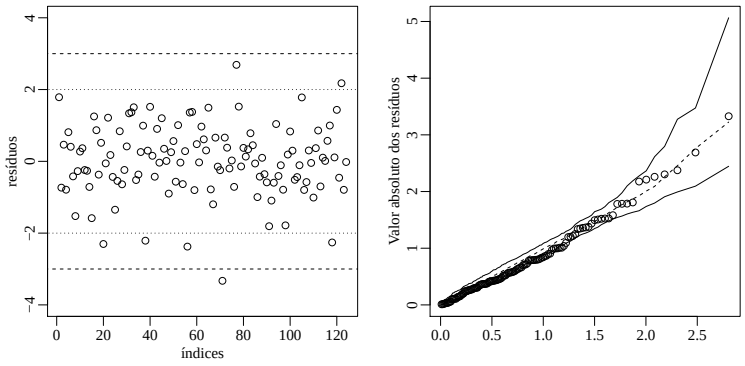
Figura 3 - Relação entre descrença religiosa e índice de inteligência.

Cabe destacar que no modelo de Cribari-Neto e Souza (2013) é modelado o parâmetro de precisão ϕ e não o parâmetro de dispersão σ. Em Cribari-Neto

e Souza (2013) apenas a variável IQ foi considerada para modelar a precisão; em nosso trabalho, além de IQ, as variáveis $MUSL$ e $logOPEN$ também se mostraram significativas no modelo da dispersão.

Comparativamente ao modelo que o teste RESET indicou como sendo o correto, ajustamos o modelo com ligações logit em ambas estruturas de regressão, evidenciado na Tabela 6. Pode-se verificar que ao utilizar a função de ligação incorreta as inferências do modelo são alteradas. Neste modelo, ao nível de 5% de significância, a variável $OPEN^2$ deve ser excluída do submodelo da média. Retirando a variável $OPEN^2$ e realizando novamente o teste RESET, com logit como ligação do modelo, o mesmo continua incorretamente especificado com $p-$valor $= 0{,}0034$.

A Figura 4 apresenta uma breve análise de diagnóstico do modelo corretamente especificado, da Tabela 5, considerando o resíduo ponderado padronizado 2 (Espinheira et al., 2008; Ferrari et al., 2011). Na Figura 4(a), nota-se que apenas um ponto está fora do intervalo $[-3,3]$ e na Figura 4(b) pode-se verificar um bom ajuste do modelo, principalmente quanto a correta distribuição dos dados. Esta análise gráfica deixa clara a correta especificação do modelo, sem indícios de má especificação das funções de ligação. Ainda, esse modelo ajustado apresenta um pseudo-R^2 (Nagelkerke, 1991) igual a 0,8105 e AIC (Akaike, 1974) igual a $-546{,}3517$.



(a) resíduos (b) envelope simulado

Figura 4 - Gráficos de diagnóstico do modelo ajustado considerando o resíduo ponderado padronizado 2.

5 Conclusões

O presente trabalho avaliou os efeitos da incorreta especificação da função de ligação do modelo da dispersão sobre as inferências no modelo de regressão beta com dispersão variável. Observou-se que a má especificação do submodelo da dispersão

causa perda de eficiência dos estimadores do submodelo da média, afeta a taxa de cobertura e o balanceamento dos valores previstos da média e da dispersão, assim como introduz viés considerável em $\hat{\sigma}_t$.

Quando são consideradas as estimativas dos parâmetros do submodelo da média, as maiores distorções da taxa de cobertura e do balanceamento se dão quando a função de ligação correta para a dispersão é a logit ou a Cauchy. Quando avaliada a capacidade de predição de σ_t os resultados numéricos evidenciaram grandes distorções em todos os cenários, tanto relativo à taxa de cobertura e balanceamento, quanto para viés relativo.

Como aplicações em dados reais são muito usuais em distintas áreas de conhecimento, considerou-se, por fim, uma aplicação em dados sobre a proporção de descrentes. Pode-se verificar que considerar apenas função de ligação logit, como é usual em trabalhos aplicados, sem considerar testes e dar devida preocupação a seleção da função de ligação adequada, pode acarretar em resultados inferenciais incorretos sobre o modelo de regressão beta com dispersão variável.

Agradecimentos

Os autores agradecem à FAPERGS, à CAPES e ao CNPq pelo auxílio financeiro recebido.

CANTERLE, D. R.; PALM, B. G.; BAYER, F. M. Effects of misspecification of the link functions in beta regression model with varying dispersion. *Rev. Bras. Biom.,* São Paulo, v.33, n.3, p.378-394, 2015.

- *ABSTRACT: The beta regression model with varying dispersion is used to model continuous data in the interval* (0,1)*, assuming beta distribution for the variable of interest. In this model, regression structures are considered for the mean and dispersion parameters involving covariates, unknown parameters and link functions. This paper addresses the problem of the misspecification in the dispersion submodel of beta regression. Coverage rates and balancing of the regression parameters in the mean submodel were evaluated by Monte Carlo simulations, considering different link functions. Coverage rate and maximum and minimum average balance of the predicted mean and dispersion values were also assessed. The misspecification of the link function of the dispersion submodel influences on the inferences of the model. An application to real data is also presented and discussed.*

- *KEYWORDS: Beta regression models with variable dispersion; confidence intervals; link function; Monte Carlo simulations.*

Referências

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* v.19, n.6, p.716-726, 1974.

ANDRADE, A. C. G. *Efeitos da especificação incorreta da função de ligação no modelo de regressão beta.* Dissertação de Mestrado, Universidade Federal de São Paulo, 2007.

BAYER, F. M., CRIBARI-NETO, F. Model selection criteria in beta regression with varying dispersion. *Communications in Statistics - Simulation and Computation,* doi: 10.1080/03610918.2014.977918, 2015.

CANTERLE, D. R., BAYER, F. Testes de especificação para a função de ligação em modelos lineares generalizados para dados binários. *Ciência e Natura,* v.37, n.1, p.1-11, 2015.

CRIBARI-NETO, F., SOUZA, T. C. Testing inference in variable dispersion beta regressions. *Journal of Statistical Computation and Simulation,* v.82, n.12, p.1827-1843, 2012.

CRIBARI-NETO, F., SOUZA, T. C. Religious belief and intelligence: Worldwide evidence. *Intelligence,* v.41, n.5, p.482-489, 2013.

ESPINHEIRA, P., FERRARI, S. L. P., CRIBARI-NETO, F. On beta regression residuals. *Journal of Applied Statistics,* v.35, n.4, p.407-419, 2008.

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

FERRARI, S. L. P., ESPINHEIRA, P. L., CRIBARI-NETO, F. Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica,* v.65, n.3, p.337-351, 2011.

FERRARI, S. L. P., PINHEIRO, E. C. Improved likelihood inference in beta regression. *Journal of Statistical Computation and Simulation,* v.81, n.4, p.431-443, 2011.

KOENKER, R., YOON, J. Parametric links for binary choice models: A fisherian-bayesian colloquy. *Journal of Econometrics,* v.152, n.2, p.120-130, 2009.

LEMONTE, J., SILVA, T. F. N. M. Estimação pontual e intervalar no modelo logístico hierárquico de dois níveis. *Revista de Matemática e Estatística,* v.24, n.2, p.147-162, 2006.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika,* v.78, n.3, p.691-692, 1991.

NEYMAN, J., PEARSON, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika,* p.17–240, 1928.

OLIVEIRA, J. S. C. *Detectando má especificação em regressão beta.* Dissertação de Mestrado, Universidade Federal de Pernanbuco, 2013.

OLIVEIRA, M. S., FERRARI, S. L. P. Inferência em um modelo de regressão beta: resultados numéricos. In: 49 REUNIÃO DA RBRAS, 2004.

PEREIRA, T. L., CRIBARI-NETO, F. Detecting model misspecification in inflated beta regressions. *Communications in Statistics - Simulation and Computation,* v.43, n.3, p.631-656, 2013.

PRESS, W., TEUKOLSKY, S., VETTERLING, W., FLANNERY, B. *Numerical recipes in C: The art of scientific computing.* Cambridge University Press, 1992.

R DEVELOPMENT CORE TEAM R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing,* Vienna, Austria, ISBN 3-900051-07-0, 2014.

RAMALHO, E., RAMALHO, J., MURTEIRA, J. Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys,* v.25, n.1, p.19-68, 2011.

RAMSEY, J. B. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society,* v.31, n.2, p.350-371, 1969.

RIGBY, R. A., STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Applied Statistics,* v.54, n.3, p.507-554, 2005.

SILVA, C. R., SOUZA, T. C. Modelagem da taxa de analfabetismo no estado da paraíba via modelo de regressão beta. *Revista Brasileira de Biometria,* v.32, n.3, p.345-359, 2014.

SIMAS, A. B., BARRETO-SOUZA, W., V., R. A. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis,* v.54, n.2, p.348-366, 2010.

SMITHSON, M., VERKUILEN, J. A better lemon squeezer Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods,* v.11, n.1, p.54-71, 2006.

SMYTH, G. K., VERBYLA, A. P. Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics,* v.10, n.6, p.695-709, 1999.

SOUZA, T. C., CRIBARI-NETO, F. Uma estimativa do impacto eleitoral do programa bolsa família. *Revista Brasileira de Biometria,* v.31, n.1, p.79-103, 2013.

WU, L., LI, H. Variable selection for joint mean and dispersion models of the inverse gaussian distribution. *Metrika,* v.75, n.6, p.795-808, 2012.

Recebido em 02.03.2015.

Aprovado após revisão em 09.07.2015.