ERROS DE ESPECIFICAÇÃO NO MODELO DE REGRESSÃO BETA COM DISPERSÃO VARIÁVEL

André Antonio de OLIVEIRA[1]
Tatiene Correia de SOUZA[1]
Saul de Azêvedo SOUZA[1]

■ RESUMO: O nosso objetivo com o presente artigo é avaliar os efeitos de erros de especificação nas inferências do modelo de regressão beta com dispersão variável. Para isto, um estudo de simulação foi realizado. Neste estudo, a variável resposta foi gerada com distribuição beta assumindo covariáveis e funções de ligação conhecidas, em sequência, o modelo foi ajustado sob a especificação correta e incorreta considerando seis tipos de erros de especificação. Avaliamos os efeitos destes erros através de taxas de rejeição e taxas de cobertura em relação a um dos parâmetros do submodelo da média e, além disso, avaliamos também o viés relativo e o erro quadrático médio em relação às estimativas das respostas médias. Verificamos através dos resultados obtidos que os erros de especificação que envolviam o preditor linear da estrutura de regressão do parâmetro de precisão apresentaram uma influência considerável nas inferências do modelo. Por fim, realizamos uma aplicação a dados reais com o objetivo de comparar na prática os efeitos de diferentes formas de especificação nas inferências do modelo de regressão beta com dispersão variável.

■ PALAVRAS-CHAVE: Modelo de regressão beta; erros de especificação; taxas de rejeição.

1 Introdução

A análise de regressão é uma das técnicas estatísticas bastante utilizadas sendo útil para investigar o comportamento de uma variável aleatória de interesse (variável dependente) quando o mesmo é influenciado por um conjunto de outras variáveis (variáveis independentes). Um dos modelos bastante utilizados em análises

[1]Universidade Federal da Paraíba - UFPB, Departamento de Estatística, CEP: 58089-900, João Pessoa, PB, Brasil. E-mail: *andreoliveira53@hotmail.com; tatiene@de.ufpb.br; saul_asouza@hotmail.com*

empíricas é o modelo de regressão normal linear. Porém, o mesmo torna-se inapropriado quando a variável resposta assume valores pertencentes a um intervalo limitado na reta, tais como taxas e proporções contínuas. Dados desta natureza usualmente se distribuem assimetricamente, não sendo adequado o uso do modelo de regressão normal linear (CRIBARI–NETO e ZEILEIS, 2010; KIESCHNICK e MCCULLOUGH, 2003).

Ferrari e Cribari–Neto (2004) propuseram um modelo de regressão que é de ampla utilidade para modelar variáveis pertencentes ao intervalo contínuo $(0, 1)$. O modelo de regressão beta proposto por estes autores assume que a variável resposta possui distribuição beta e que sua média é relacionada a um preditor linear por meio de uma função de ligação, preditor este que envolve covariáveis e parâmetros de regressão desconhecidos. Este modelo também é indexado por um parâmetro de dispersão, que neste caso, é constante ao longo das observações. Contudo, Simas *et al.* (2010) apresentaram uma extensão do modelo proposto por Ferrari e Cribari–Neto (2004), denominado modelo de regressão beta com dispersão variável. Nesta abordagem, o parâmetro de dispersão varia ao longo das observações, sendo modelado também por uma estrutura de regressão que contém covariáveis, parâmetros desconhecidos e uma função de ligação. Segundo Espinheira *et al.* (2008*b*), a classe de modelos de regressão beta é similar em muitos aspectos à classe dos modelos lineares generalizados (MCCULLAGH e NELDER, 1989).

Diversas aplicações do modelo de regressão beta podem ser encontradas na literatura. Oliveira e Souza (2016), por exemplo, utilizaram este modelo para investigar a proporção de crianças obesas beneficiadas pelo Programa Bolsa Família nas regiões do Brasil. Almeida Junior e Souza (2015) avaliaram o impacto exercido pelo Programa Bolsa Família nas eleições presidenciais do ano de 2010, enquanto que Silva e Souza (2014) tiveram por finalidade modelar a taxa de analfabetismo nos municípios do estado da Paraíba neste mesmo ano. Sant'Anna e Caten (2010) modelaram a fração ou proporção de itens não conformes às especificações de um processo industrial com enfoque no modelo de regressão beta e no modelo linear generalizado, por outro lado, Pinto *et al.* (2011) fizeram uso destes mesmos modelos de regressão para um estudo relacionado à infartos em pacientes. Outros exemplos de aplicações podem ser encontradas em Souza e Cribari–Neto (2015) e Cribari–Neto e Pereira (2013).

Ao se realizar qualquer análise de regressão não é possível saber de fato se o modelo estimado retrata adequadamente a realidade do fenômeno em estudo. Segundo Pereira e Cribari–Neto (2014), caso a especificação do modelo escolhido esteja incorreta, inferências imprecisas podem vir a ocorrer. No modelo de regressão beta com dispersão variável a escolha das funções de ligação e das variáveis independentes são processos tipicamente necessários no ajuste de um determinado modelo, e que podem ser auxiliados por conclusões de estudos anteriores. Porém, em termos práticos, é comum se cometer erros de especificação neste processo.

Neste contexto, alguns autores estudaram diferentes formas de especificações na classe de modelos de regressão beta. Bayer e Cribari–Neto (2015) exploraram

o tema de seleção das covariáveis importantes nas estruturas de regressão dos submodelos da média e da dispersão. Andrade (2007) realizou um extenso estudo de simulação com o objetivo de avaliar o impacto da especificação incorreta da função de ligação da média e comparou, através de uma aplicação prática, os resultados obtidos através do uso de diferentes funções de ligação. Lima (2007) propôs um teste de erro de especificação para modelos de regressão beta baseado no teste *RESET* (RAMSEY, 1969), e concluiu através de simulações que o mesmo é útil para detecção do uso de função de ligação incorreta bem como de não-linearidades no preditor linear. Canterle *et al.* (2015) abordaram o problema da má especificação na função de ligação do submodelo da dispersão e verificaram que a incorreta especificação desta função de ligação tem uma influência considerável nas inferências do modelo, incluindo implicações diretas na eficiência dos estimadores dos parâmetros da média.

Loose *et al.* (2014) abordaram o desempenho dos estimadores pontuais e intervalares no modelo de regressão beta com dispersão variável e, através de simulações de Monte Carlo, confirmaram a consistência destes estimadores. Contudo, observaram que os estimadores que modelam a precisão (quanto maior a precisão menor a dispersão) são consideravelmente mais viesados do que os que modelam a média, indicando uma necessidade de maior atenção na modelagem da estrutura de regressão deste parâmetro. Considerando a dificuldade em se modelar esta estrutura e que na prática é comum se cometer erros de especificação, Cribari–Neto e Souza (2012) propuseram uma nova abordagem em modelos de regressão beta como forma de solucionar estes problemas. Esta abordagem é baseada em estimadores do tipo sanduíche para casos em que a estrutura de regressão para o parâmetro de dispersão é negligenciada. Os autores concluíram que as inferências considerando esta metodologia são precisas mesmo sob dispersão variável, o que indica uma possível solução para os erros de especificação que usualmente são cometidos nesta estrutura.

Neste contexto de erros de especificação e tendo como base alguns trabalhos citados anteriormente, a exemplo de Andrade (2007) e Canterle *et al.* (2015), o nosso objetivo aqui é avaliar o efeito destes erros nas inferências do modelo de regressão beta com dispersão variável. Um estudo de simulação considerando diferentes cenários foi realizado com este propósito. Nestas simulações, a variável resposta foi gerada com distribuição beta assumindo covariáveis e funções de ligação conhecidas, o modelo foi então ajustado considerando a especificação correta e incorreta. Em particular, seis tipos de erros de especificação foram avaliados, englobando tanto erros nos preditores quanto nas funções de ligação das duas estruturas de regressão. Para avaliar o efeito destes erros, consideramos as taxas de rejeição e as taxas de cobertura em relação a um dos parâmetros do submodelo da média (μ). Computamos ainda algumas medidas considerando as estimativas para as respostas médias, a saber: o viés relativo médio e o erro quadrático médio. Por fim, realizamos uma aplicação a dados reais.

O presente artigo encontra-se dividido em cinco seções. A Seção 2 apresenta o modelo de regressão beta com dispersão variável. Os resultados numéricos e as

discussões são apresentados na Seção 3. Na Seção 4 uma aplicação a dados reais é realizada. Por último, na Seção 5 são apresentadas as conclusões e considerações finais.

2 Modelo de regressão Beta

O modelo de regressão beta, introduzido por Ferrari e Cribari–Neto (2004), é comumente utilizado para modelar variáveis que assumem valores no intervalo $(0, 1)$, a exemplo de taxas e proporções. Almeida Junior e Souza (2015), Oliveira e Souza (2016), Smithson e Verkuilen (2006) e Souza e Cribari–Neto (2015) utilizaram modelos de regressão para a situação em que a variável resposta segue distribuição beta. Em tais modelos assume-se que a resposta média é relacionada a um preditor linear através de uma função de ligação, preditor este que envolve covariáveis e parâmetros de regressão desconhecidos. Estes modelos também são indexados por um parâmetro de dispersão que em certas ocasiões podem variar ao longo das observações (CRIBARI–NETO e SOUZA, 2012, 2013; ESPINHEIRA *et al.*, 2008a, 2008b; SILVA e SOUZA, 2014; SIMAS *et al.*, 2010).

Para a definição do modelo de regressão beta, Ferrari e Cribari–Neto (2004) sugerem uma parametrização da distribuição beta em termos de sua média e um parâmetro de precisão. Com essa parametrização a função densidade da distribuição beta pode ser reescrita como

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

em que $0 < \mu < 1$ e $\phi > 0$. Aqui, $\mathrm{E}(y) = \mu$ e $\mathrm{var}(y) = \frac{V(\mu)}{1+\phi}$, sendo $V(\mu) = \mu(1-\mu)$, a 'função variância', μ é a média da variável resposta e ϕ pode ser interpretado como o parâmetro de precisão no sentido que, para um valor fixo de μ, quanto maior o valor de ϕ, menor a variância de y. O parâmetro de dispersão é obtido considerando $\sigma^2 = \frac{1}{1+\phi_t}$.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, em que cada y_t, $t = 1, \ldots, n$, segue a densidade apresentada em (1) com média μ_t e parâmetro de precisão ϕ_t sendo desconhecidos. O modelo de regressão beta com precisão constante assume que uma função da média μ_t pode ser igualada ao preditor linear η_t, sendo esta estrutura definida por $g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t$, em que $\beta = (\beta_1, \ldots, \beta_k)^\top$ é um vetor de parâmetros de regressão desconhecidos ($\beta \in \mathbb{R}^k$), $x_{t1}, \ldots, x_{tk}$ são observações de k covariáveis e $g(\cdot)$ é denominada função de ligação. Portanto, $\mu_t = g^{-1}(\eta_t)$ e $\mathrm{var}(y_t) = \mu_t(1-\mu_t)/(1+\phi)$, para $t = 1, \ldots, n$.

O modelo de regressão beta proposto por Ferrari e Cribari–Neto (2004) considera o parâmetro de precisão constante ao longo das observações. Porém, admitimos como em Simas *et al.* (2010) que o parâmetro de precisão é variável, sendo modelado através de uma estrutura de regressão que contém covariáveis, parâmetros de regressão desconhecidos e uma função de ligação, sendo esta estrutura definida por $h(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = \vartheta_t$, em que $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ é um vetor de

parâmetros desconhecidos, $z_{t1}, \ldots, z_{tq}$ são observações de q covariáveis ($k + q<n$) assumidas fixas e conhecidas, ϑ_t é o preditor linear e $h(\cdot)$ é uma função de ligação. Aqui, $\phi_t = h^{-1}(\vartheta_t)$, em que $t = 1, \ldots, n$.

As estimativas dos parâmetros são obtidas maximizando numericamente a função de log-verossimilhança através de um algoritmo de maximização não-linear. Usualmente, utiliza-se o método quasi-Newton BFGS (PRESS *et al.*, 1992). A distribuição dos estimadores de máxima verossimilhança de β e γ, ditos $\widehat{\beta}$ e $\widehat{\gamma}$, é aproximadamente normal em grandes amostras. Esta aproximação pode ser usada na construção de intervalos de confiança e testes de hipóteses. Para maiores detalhes inferenciais e matriciais do vetor escore e da matriz de informação de Fisher, ver Simas *et al.* (2010).

2.1 Teste de hipóteses e intervalos de confiança

De acordo com Ferrari e Cribari–Neto (2004) inferências em grandes amostras no modelo de regressão beta podem ser realizadas utilizando o teste de Wald (WALD, 1943). A estatística de teste para testar a hipótese nula $\mathcal{H}_0 : \beta_1 = \beta_1^{(0)}$ é dada por:

$$\omega = (\widehat{\beta}_1 - \beta_1^{(0)})^\top (\widehat{K}_{11}^{\beta\beta})^{-1} (\widehat{\beta}_1 - \beta_1^{(0)}),$$

em que $\widehat{K}_{11}^{\beta\beta}$ é igual $K_{11}^{\beta\beta}$ (matriz $m \times m$ correspondente a β obtida da inversa da matriz de informação de Fisher K^{-1}) avaliado no estimador de máxima verossimilhança irrestrito, e $\widehat{\beta}_1$ é o estimador de máxima verosimilhança de β_1. Sob fracas condições de regularidade e sob $\mathcal{H}_0$, $\omega \xrightarrow{D} \chi^2_m$. Em particular, para testar a significância do i-ésimo parâmetro de regressão (β_i), $i = 1, \ldots, k$, pode-se utilizar a raiz quadrada da estatística de Wald (teste z), isto é, $\widehat{\beta}_i / ep(\widehat{\beta}_i)$, onde $ep(\widehat{\beta}_i)$ é o erro padrão assintótico do estimador de máxima verossimilhança de $\widehat{\beta}_i$ obtido da inversa da matriz de informação de Fisher avaliada nas estimativas de máxima verossimilhança. A distribuição nula restrita da estatística de teste é normal padrão. Maiores detalhes sobre testes de hipóteses no modelo de regressão beta podem ser vistos em Ferrari e Cribari–Neto (2004) e Cribari–Neto e Queiroz (2014).

Um intervalo de confiança $(1 - \alpha) \times 100\%$ para os parâmetros dos modelos, sendo $\theta = (\beta^\top, \gamma^\top)^\top$ o vetor de parâmetros, é dado por:

$$\widehat{\theta}_i \pm \Phi^{-1}(1 - \alpha/2) ep(\widehat{\theta}_i),$$

em que Φ^{-1} é a função de distribuição acumulada de uma variável aleatória normal padrão, $ep(\widehat{\theta}_i)$ é o erro padrão para $\widehat{\theta}_i$ e α é o nível nominal do intervalo de confiança.

2.2 Funções de ligação

As funções $g(\cdot)$ e $h(\cdot)$ são conhecidas como funções de ligação e existem muitas possibilidades para suas escolhas. Considerando o parâmetro da média temos que, $g(\mu_t) = \eta_t$, $t = 1, \ldots, n$, em que $g(\cdot)$ é estritamente monótona e duas vezes

diferenciável, com domínio em $(0,1)$ e imagem em $\mathbb{R}$. Portanto, $\mu_t = g^{-1}(\eta_t)$. Alguns exemplos destas funções de ligação são:

- logit: $g(\mu_t) = \log(\frac{\mu_t}{1-\mu_t})$; $\mu_t = \frac{\exp(\eta_t)}{1+\exp(\eta_t)}$;
- probit: $g(\mu_t) = \Phi^{-1}(\mu_t)$; em que $\Phi^{-1}(\cdot)$ é a função de distribuição acumulada de uma variável normal padrão; $\mu_t = \Phi(\eta_t)$;
- cloglog: $g(\mu_t) = \log\{-\log(1-\mu_t)\}$; $\mu_t = 1 - \exp\{-\exp(\eta_t)\}$;
- loglog: $g(\mu_t) = -\log\{-\log(\mu_t)\}$; $\mu_t = \exp\{-\exp(-\eta_t)\}$.

em que $t = 1, \ldots, n$. A Figura 1 apresenta o comportamento de η como função de μ para as funções de ligação citadas acima. É possível perceber, por exemplo, que as funções logit e probit tem comportamento parecido, que a função cloglog tem comportamento similar à logit para valores de μ próximos de 0 e a função de ligação loglog tem comportamento similar à logit para valores de μ próximos de 1.

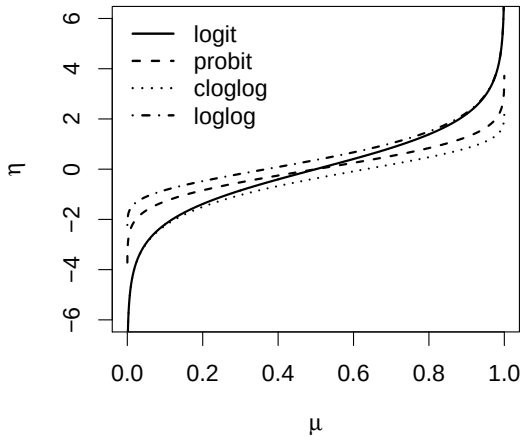


Figura 1 - Gráfico de funções de ligação para a média.

Para o parâmetro de precisão temos que, $h(\phi_t) = \vartheta_t$, em que $h(\cdot)$ é uma função estritamente monótona e duas vezes diferenciável que mapeia os pontos positivos da reta. Portanto, $\phi_t = h^{-1}(\vartheta_t)$. Alguns exemplos destas funções são:

- log: $h(\phi_t) = \log(\phi_t)$; $\phi_t = \exp\{\vartheta_t\}$;
- sqrt: $h(\phi_t) = \sqrt{\phi_t}$; $\phi_t = \vartheta_t^2$;

- identity: $h(\phi_t) = \phi_t$; $\phi_t = \vartheta_t$.

em que $t = 1, \ldots, n$. Para maiores detalhes sobre as funções de ligação ver McCullagh e Nelder (1989).

3 Avaliação numérica

Através de simulações de Monte Carlo nós avaliamos, em amostras de tamanho finito, o efeito de diferentes erros de especificação no modelo de regressão beta com dispersão variável. Os erros de especificação cometidos estão apresentados na Tabela 1 e foram avaliados em diferentes cenários, considerando tamanhos amostrais e intervalos de médias distintos. A implementação computacional foi desenvolvida no *software* estatístico `R` (`R` CORE TEAM, 2017; KLEIBER e ZEILEIS, 2008) utilizando o pacote betareg (CRIBARI–NETO e ZEILEIS, 2010).

O número de réplicas de Monte Carlo foi fixado em 10.000. Para cada réplica de Monte Carlo foram geradas amostras aleatórias da variável aleatória y_t, $t = 1, \ldots, n$, com função de densidade dada em (1), parâmetro de média definido por $\mu_t = g^{-1}(\eta_t)$ e parâmetro de precisão definido por $\phi_t = h^{-1}(\vartheta_t)$, em que:

$$\eta_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t},$$
$$\vartheta_t = \gamma_0 + \gamma_1 z_{1t},$$

para $t = 1, \ldots, n$. As variáveis independentes foram geradas a partir da distribuição uniforme $(0, 1)$. Na geração dos dados, consideramos o modelo corretamente especificado (MCE) definido pela estrutura apresentada acima, e consideramos ainda $g(\cdot)$ a função de ligação logit e $h(\cdot)$ a função de ligação sqrt. Medimos a não-constância da precisão dos dados através da quantidade $\lambda = \frac{\max(\phi_t)}{\min(\phi_t)}$, para $t = 1, \ldots, n$. Note que $\lambda = 1$ indica que a precisão é constante para todas as observações. No nosso estudo $\lambda \approx 13$.

Com o objetivo de avaliar os efeitos dos erros de especificação em diferentes cenários, consideramos três intervalos para as médias na geração dos dados. No primeiro cenário, utilizamos $\beta_0 = -1.9$, $\beta_1 = 1.5$ e $\beta_2 = 0.0$ como verdadeiros valores dos parâmetros, que conduziram a valores de médias próximos à 0, mais precisamente $\mu_t \in [0.1374; 0.3764]$. No segundo cenário, os valores para os parâmetros foram $\beta_0 = -0.5$, $\beta_1 = 1.2$ e $\beta_2 = 0.0$, que conduziram a $\mu_t \in [0.3894; 0.6493]$, isto é, médias próximas à 0.5. No terceiro cenário, utilizamos $\beta_0 = 2.3$, $\beta_1 = -1.8$ e $\beta_2 = 0.0$, que neste caso conduziu a $\mu_t \in [0.6515; 0.9024]$, ou seja, valores de médias próximas à 1. No caso da precisão, os verdadeiros valores para os parâmetros foram $\gamma_0 = 1.0$ e $\gamma_1 = 7.9$, que produziram valores de ϕ_t no intervalo $[5.8420; 76.6020]$, o que corresponde a um cenário de baixa precisão, ou ainda, de alta dispersão.

Consideramos ainda três diferentes tamanhos amostrais, $n = 25, 50, 100$, sendo que são geradas 25 observações das variáveis x_1, x_2, z_1 e z_2, que são replicadas duas

Tabela 1 - Descrição dos cenários considerados no estudo de simulação de Monte Carlo

Especificação dos modelos	Geração	
	μ	ϕ
Modelo Corretamente Especificado - MCE	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\text{sqrt}(\phi_t) = \gamma_0 + \gamma_1 z_{1t}$
Especificação dos modelos	Estimação	
	μ	ϕ
Erro no preditor linear da média e nas funções de ligação da média e precisão - E1	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}^2$	$\log(\phi_t) = \gamma_0 + \gamma_1 z_{1t}$
Erro no preditor linear da precisão e na função de ligação da precisão - E2	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\log(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Erros nos preditores lineares da média e da precisão e nas funções de ligação da média e precisão - E3	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}^2$	$\log(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Erro no preditor linear da precisão e na função de ligação da média - E4	$\text{cloglog}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\text{sqrt}(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Erros nos preditores lineares da média e da precisão - E5	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}^2$	$\text{sqrt}(\phi_t) = \gamma_0 + \gamma_1 z_{2t}$
Precisão fixa - E6	$\text{logit}(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}$	$\text{sqrt}(\phi_t) = \gamma_0$

e quatro vezes, respectivamente, para os tamanhos amostrais $n = 50, 100$. Este procedimento de replicação de valores assegura que o grau de heterogeneidade na precisão dos dados mantenha-se constante à medida em que se aumenta o tamanho amostral.

Na avaliação dos efeitos dos erros de especificação nas inferências do modelo de regressão beta foram computadas as taxas de rejeição sob a hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ e os intervalos de confiança com nível nominal de confiança de 95% para o parâmetro β_2. A partir de 10.000 intervalos de confiança foram obtidas as taxas de cobertura para β_2. A taxa de cobertura representa a proporção de vezes em que o intervalo de confiança conteve o parâmetro, sendo esperada uma taxa em torno de 95%. Em relação às estimativas das médias μ_t, foram avaliados os vieses relativos médios (VRm) e o erro quadrático médio (EQM), sendo definidos por:

$$VRm = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{\mu_t} - \mu_t|}{\mu_t},$$

$$EQM = \frac{1}{n}\sum_{t=1}^{n}E[(\hat{\mu_t} - \mu_t)^2],$$

em que $t = 1, \ldots, n$, $\hat{\mu_t}$ é a estimativa para as médias considerando as n observações e E denota a operação de valor esperado sob a quantidade $(\hat{\mu_t} - \mu_t)^2$. Para essas medidas espera-se valores próximos de zero com o aumento do tamanho amostral.

Vale salientar aqui que estamos considerando o valor médio destas medidas em relação aos 10.000 resultados que serão gerados pela simulação de Monte Carlo.

Primeiramente, nosso interesse consiste em avaliar as taxas de rejeição (tamanho dos testes z) sob a hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ versus $\mathcal{H}_1 : \beta_2 \neq 0$ considerando seis tipos de erros de especificação (ver Tabela 1). Na Tabela 2 são apresentados os resultados das taxas de rejeição aos níveis nominais de 10%, 5% e 1% considerando diferentes intervalos para as médias μ_t. As principais conclusões serão resumidas a seguir. Em primeiro lugar, considerando o modelo corretamente especificado, é possível observar que as taxas de rejeição estão acima dos níveis nominais considerados para os três intervalos das médias e para os diferentes tamanhos amostrais. Vale salientar que as mesmas foram ainda mais distantes dos níveis nominais para valores de médias próximas de 0. Como ilustração, considerando o cenário em que as médias estão variando no intervalo $[0.1374; 0.3764]$, $n = 50$ e $\alpha = 10\%$, a taxa de rejeição é de 13.07%. No cenário 2, de médias entre $[0.3894; 0.6493]$, essa mesma taxa é de 12.26%. Já para o cenário 3, cujo intervalo de médias é entre $[0.6515; 0.9024]$ a taxa é de 12.65%.

Segundo, as taxas de rejeição tendem a ser maiores quando há erro no preditor da precisão (E2, E3, E4 e E5) comparado ao ajuste em que não há esse erro (E1) ou quando o modelo é estimado com precisão fixa (E6). Exemplificando, considere o cenário 2, com médias variando próximas à 0.5, $n = 50$ e $\alpha = 5\%$, a taxa de rejeição do modelo em que há erros nas funções de ligação da média e da precisão e no preditor da média (E1) é de 6.71%, enquanto que no modelo em que há esses mesmos erros acrescido ao erro no preditor da precisão (E3), essa taxa é de 10.24%. Considerando o ajuste estimado com precisão fixa (E6), a taxa de rejeição associada é de 7.74%.

Terceiro, no caso em que o modelo foi estimado cometendo erros no preditor da precisão e em uma das funções de ligação, nossos resultados mostraram que no geral, é mais grave se cometer erros no preditor da precisão e na função de ligação da média do que no preditor e na função de ligação do parâmetro de precisão, considerando valores de médias variando próximos de 0 e de 0.5 (cenários 1 e 2). Já para valores de médias próximos de 1 (cenário 3), os resultados se inverteram, as taxas maiores se associaram ao erro no preditor e na função de ligação do parâmetro de precisão. Como exemplo considere o cenário 1, $n = 100$ e $\alpha = 5\%$, a taxa de rejeição para o modelo em que se cometem erros no preditor e na função de ligação da precisão (E2) é de 9.07%, enquanto que para o modelo estimado com erros no preditor da precisão e na função de ligação da média (E4), essa mesma taxa é de 9.73%. No cenário 3, considerando o mesmo tamanho amostral e nível de significância, essas taxas foram de 9.13% e 8.12%, respectivamente.

Quarto, considerando os erros de especificação cometidos, o modelo estimado que no geral apresentou taxas mais distantes dos níveis nominais, para os cenários 1 e 2, foi o modelo com erros nas funções de ligação e nos preditores dos parâmetros da média e da precisão (E3). Por exemplo, para valores de médias próximos de 0.5, $n = 50$ e $\alpha = 5\%$, o tamanho do teste baseado no modelo estimado com os erros de especificação descritos acima rejeita a hipótese nula 10.24% das vezes, o

Rev. Bras. Biom., Lavras, v.36, n.3, p.601-624, 2018 - doi: 10.28951/rbb.v36i3.245

Tabela 2 - Taxas de rejeição sob $\mathcal{H}_0 : \beta_2 = 0$

Cenário 1	$\mu_t \in [0.1374; 0.3764]$								
	$n = 25$			$n = 50$			$n = 100$		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
Modelo corretamente especificado - MCE	16.10	9.70	5.35	13.07	7.54	3.40	11.52	6.17	2.75
Erro no pred. linear da média e nas funções de lig. - E1	17.37	10.58	6.13	13.12	7.31	3.49	11.18	6.19	2.67
Erro no pred. linear e na função de lig. da precisão - E2	21.15	13.50	7.76	16.89	10.10	5.22	15.05	9.07	4.66
Erros nos pred. lineares e nas funções de lig. - E3	22.07	14.83	8.89	17.69	11.01	6.05	16.98	10.40	5.38
Erro no pred. linear da precisão e na função de lig. da média - E4	21.18	13.68	7.77	17.15	10.42	5.44	16.06	9.73	4.91
Erros nos pred. lineares da média e da precisão - E5	21.90	14.18	8.54	17.24	10.69	5.82	16.22	9.99	5.09
Precisão Fixa - E6	17.14	10.00	5.21	14.91	8.53	4.12	13.70	7.93	3.76
Cenário 2	$\mu_t \in [0.3894; 0.6493]$								
	$n = 25$			$n = 50$			$n = 100$		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
Modelo corretamente especificado - MCE	15.88	9.56	5.11	12.26	6.73	3.12	11.36	6.32	2.76
Erro no pred. linear da média e nas funções de lig. - E1	16.13	10.09	5.58	12.00	6.71	3.02	11.26	6.38	2.59
Erro no pred. linear e na função de lig. da precisão - E2	20.91	13.90	8.12	15.42	9.19	4.64	14.21	8.16	3.97
Erros nos pred. lineares e nas funções de lig. - E3	21.19	14.31	8.53	16.04	10.24	5.34	15.31	8.98	4.65
Erro no pred. linear da precisão e na função de lig. da média - E4	20.73	13.72	8.00	15.80	9.39	4.78	14.65	8.41	4.24
Erros nos pred. lineares da média e da precisão - E5	20.32	13.51	7.98	15.80	9.67	5.03	14.52	8.62	4.38
Precisão Fixa - E6	16.06	10.01	5.11	13.67	7.74	3.69	13.40	7.74	3.58
Cenário 3	$\mu_t \in [0.6515; 0.9024]$								
	$n = 25$			$n = 50$			$n = 100$		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
Modelo corretamente especificado - MCE	15.87	9.36	5.22	12.65	7.13	3.61	10.70	5.87	2.51
Erro no pred. linear da média e nas funções de lig. - E1	17.10	10.62	5.97	13.05	7.36	3.70	11.76	6.03	2.70
Erro no pred. linear e na função de lig. da precisão - E2	20.40	13.37	7.96	17.38	10.62	5.26	15.30	9.13	4.50
Erros nos pred. lineares e nas funções de lig. - E3	20.88	13.26	7.69	16.81	9.96	5.10	14.67	8.45	4.14
Erro no pred. linear da precisão e na função de lig. da média - E4	19.84	12.70	7.47	16.44	9.92	4.80	13.99	8.12	3.89
Erros nos pred. lineares da média e da precisão - E5	21.46	14.29	8.60	18.40	11.66	6.12	16.43	9.69	5.13
Precisão Fixa - E6	16.47	10.06	5.52	15.21	8.93	4.28	13.66	7.91	3.84

que vale ressaltar que é mais do que o dobro do nível nominal considerado. Em contrapartida, para o cenário 3, o modelo com erros nos preditores da média e da precisão (E5) apresentou as maiores taxas de rejeição comparadas aos demais modelos estimados.

Quinto, as taxas de rejeição do modelo estimado com precisão fixa (E6), considerando os tamanhos amostrais $n = 50$ e $n = 100$, tenderam a ser maiores do que as taxas do modelo estimado com erros nas funções de ligação da média e da precisão e no preditor linear da média (E1). Isso significa que, para tamanhos amostrais maiores omitir a estrutura de regressão do parâmetro de precisão pode ser mais grave do que errar na especificação da média e na função de ligação do parâmetro de precisão, resultado este que não se verifica para amostras pequenas ($n = 25$). Para ilustrar essa conclusão considere médias variando próximas de 1 e $\alpha = 10\%$, as taxas de rejeição do modelo estimado com precisão fixa considerando os três tamanhos amostrais foram de 16.47%, 15.21% e 13.66%, enquanto que no modelo com erros nas funções de ligação e no preditor da média, essas taxas foram de 17.10%, 13.05 e 11.76%, respectivamente.

Na Tabela 3 encontram-se os resultados numéricos da avaliação das taxas de cobertura para o parâmetro β_2. Nota-se que os resultados estão dentro do esperado, pois as taxas se aproximam de 95% com o aumento do tamanho amostral, considerando os erros de especificação cometidos e os diferentes intervalos para as médias. Com base nesses resultados, temos que o modelo estimado com mais erros de especificação (E3) apresentou as taxas de cobertura mais distantes de 95% para os cenários 1 e 2, enquanto que para o cenário 3, o modelo estimado com erros nos preditores da média e da precisão (E5) apresentou as menores taxas de cobertura comparado aos demais modelos.

Assim como ocorreram com as taxas de rejeição, nos modelos em que há o erro no preditor da precisão (E2, E3, E4 e E5) há uma tendência a apresentar resultados mais distantes do esperado comparado ao caso em que não há esse erro (E1) ou quando o modelo é estimado desconsiderando a estrutura de regressão para o parâmetro de precisão (E6). Exemplificando, considere valores de médias próximos de 0 e $n = 50$, a taxa de cobertura para o parâmetro β_2 no modelo com erros nas funções de ligação e no preditor da média (E1) é de 92.69%, enquanto que no modelo estimado com esses mesmos erros acrescido ao erro no preditor da precisão (E3), a taxa é de 88.99%. Já para o modelo estimado com precisão fixa (E6), a mesma foi de 91.47%.

As taxas de cobertura para o modelo estimado com erros no preditor da precisão e na função de ligação da média (E4) tenderam a ser mais distantes do valor esperado do que no modelo estimado com erros no preditor e na função de ligação do parâmetro precisão (E2) considerando os cenários 1 e 2. Resultado que reforça o fato de que, errar no preditor da precisão e na função de ligação da média é mais grave do que errar apenas na especificação do parâmetro de precisão (função de ligação e preditor linear) considerando valores de médias variando até próximos de 0.65.

Confrontando os resultados obtidos em relação apenas aos diferentes intervalos

Tabela 3 - Resultado das taxas de cobertura para β_2

Cenário 1	$\mu_t \in [0.1374; 0.3764]$		
	$n = 25$	$n = 50$	$n = 100$
Modelo corretamente especificado - MCE	90.30	92.46	93.83
Erro no pred. linear da média e nas funções de lig. - E1	89.42	92.69	93.81
Erro no pred. linear e na função de lig. da precisão - E2	86.50	89.90	90.93
Erros nos pred. lineares e nas funções de lig. - E3	85.17	88.99	89.60
Erro no pred. linear da precisão e na função de lig. da média - E4	86.32	89.58	90.27
Erros nos pred. lineares da média e da precisão - E5	85.82	89.31	90.01
Precisão Fixa - E6	90.00	91.47	92.07
Cenário 2	$\mu_t \in [0.3894; 0.6493]$		
	$n = 25$	$n = 50$	$n = 100$
Modelo corretamente especificado - MCE	90.44	93.27	93.68
Erro no pred. linear da média e nas funções de lig. - E1	89.91	92.29	93.62
Erro no pred. linear e na função de lig. da precisão - E2	86.10	90.81	91.84
Erros nos pred. lineares e nas funções de lig. - E3	85.69	89.76	91.02
Erro no pred. linear da precisão e na função de lig. da média - E4	86.28	90.61	91.59
Erros nos pred. lineares da média e da precisão - E5	86.49	90.33	91.38
Precisão Fixa - E6	89.99	92.26	92.26
Cenário 3	$\mu_t \in [0.6515; 0.9024]$		
	$n = 25$	$n = 50$	$n = 100$
Modelo corretamente especificado - MCE	90.64	92.87	94.13
Erro no pred. linear da média e nas funções de lig. - E1	89.38	92.34	93.97
Erro no pred. linear e na função de lig. da precisão - E2	86.63	89.38	90.87
Erros nos pred. lineares e nas funções de lig. - E3	86.74	90.04	91.55
Erro no pred. linear da precisão e na função de lig. da média - E4	87.30	90.08	91.88
Erros nos pred. lineares da média e da precisão - E5	85.71	88.34	90.31
Precisão Fixa - E6	89.94	91.07	92.09

de médias considerados, temos que em geral, as taxas de cobertura tenderam a ser maiores para os cenários com médias variando próximos de 0.5 e de 1. Para ilustrar este resultado, considere o tamanho amostral $n = 100$ e caso em que se cometem erros nos preditores lineares e nas funções de ligação dos parâmetros da média e da precisão (E3), as taxas de cobertura considerando os cenários 1, 2 e 3, foram 89.60%, 91.02% e 91.55%, respectivamente.

A Tabela 4 apresenta os vieses relativos médios e o erro quadrático médio das estimativas das médias. Através de sua análise, pode-se observar que em todos os casos, como já era esperado devido a propriedade de consistência dos estimadores de máxima verossimilhança, com o aumento do tamanho amostral os valores tenderam a zero. Além do mais, o modelo estimado corretamente no geral apresentou os menores valores para essas medidas.

Dos intervalos para as médias utilizados na geração dos dados, o cenário 3, de médias próximas de 1, apresentou menores valores comparado aos outros dois cenários. Mais uma vez, modelos estimados com erro no preditor linear do parâmetro da precisão apresentaram resultados mais distantes do esperado. Vale destacar, que no caso do viés relativo médio e do erro quadrático médio, o

Tabela 4 - Vieses Relativos Médios (VRm) e Erro Quadrático Médio (EQM) para o estimador das médias

Cenário 1	$\mu_t \in [0.1374; 0.3764]$					
	VRm			EQM		
	$n = 25$	$n = 50$	$n = 100$	$n = 25$	$n = 50$	$n = 100$
Modelo corretamente especificado - MCE	0.0839	0.0589	0.0413	0.0128	0.0089	0.0070
Erro no pred. linear da média e nas funções de lig. - E1	0.0857	0.0597	0.0418	0.0132	0.0091	0.0071
Erro no pred. linear e na função de lig. da precisão - E2	0.1064	0.0739	0.0522	0.0173	0.0110	0.0081
Erros nos pred. lineares e nas funções de lig. - E3	0.1068	0.0743	0.0525	0.0175	0.0111	0.0082
Erro no pred. linear da precisão e na função de lig. da média - E4	0.1061	0.0740	0.0524	0.0173	0.0111	0.0081
Erros nos pred. lineares da média e da precisão - E5	0.1067	0.0743	0.0524	0.0174	0.0111	0.0081
Precisão Fixa - E6	0.1052	0.0750	0.0532	0.0172	0.0113	0.0083
Cenário 2	$\mu_t \in [0.3894; 0.6493]$					
	VRm			EQM		
	$n = 25$	$n = 50$	$n = 100$	$n = 25$	$n = 50$	$n = 100$
Modelo corretamente especificado - MCE	0.0450	0.0313	0.0220	0.0093	0.0078	0.0071
Erro no pred. linear da média e nas funções de lig. - E1	0.0459	0.0318	0.0224	0.0094	0.0079	0.0071
Erro no pred. linear e na função de lig. da precisão - E2	0.0570	0.0390	0.0276	0.0109	0.0085	0.0075
Erros nos pred. lineares e nas funções de lig. - E3	0.0573	0.0394	0.0281	0.0110	0.0086	0.0075
Erro no pred. linear da precisão e na função de lig. da média - E4	0.0570	0.0393	0.0280	0.0109	0.0086	0.0075
Erros nos pred. lineares da média e da precisão - E5	0.0570	0.0392	0.0278	0.0109	0.0086	0.0075
Precisão Fixa - E6	0.0565	0.0397	0.0284	0.0108	0.0086	0.0075
Cenário 3	$\mu_t \in [0.6515; 0.9024]$					
	VRm			EQM		
	$n = 25$	$n = 50$	$n = 100$	$n = 25$	$n = 50$	$n = 100$
Modelo corretamente especificado - MCE	0.0224	0.0156	0.0109	0.0067	0.0062	0.0059
Erro no pred. linear da média e nas funções de lig. - E1	0.0228	0.0162	0.0117	0.0065	0.0059	0.0057
Erro no pred. linear e na função de lig. da precisão - E2	0.0290	0.0203	0.0142	0.0072	0.0064	0.0060
Erros nos pred. lineares e nas funções de lig. - E3	0.0286	0.0204	0.0147	0.0069	0.0061	0.0058
Erro no pred. linear da precisão e na função de lig. da média - E4	0.0284	0.0203	0.0146	0.0069	0.0061	0.0057
Erros nos pred. lineares da média e da precisão - E5	0.0291	0.0205	0.0143	0.0073	0.0064	0.0060
Precisão Fixa - E6	0.0284	0.0203	0.0143	0.0074	0.0066	0.0061

modelo estimado desconsiderando a estrutura de regressão do parâmetro de precisão apresentou em alguns casos, os resultados mais distantes de zero comparado aos demais modelos aqui estudados. Exemplificando este último resultado, considere $n = 50$ e médias variando próximas de 0, o viés relativo médio e o erro quadrático médio do modelo estimado com precisão fixa (E6) foram de 0.0750 e 0.0113 respectivamente, enquanto que no modelo estimado com erros no preditor linear da média e nas funções de ligação da média e da precisão (E1), por exemplo, estes valores foram de 0.0597 e 0.0091, respectivamente.

Deste modo, através dos resultados das simulações verificamos que os erros de especificação que envolviam o preditor linear da estrutura de regressão do parâmetro de precisão tiveram uma influência considerável nas inferências do modelo de regressão beta com dispersão variável. De acordo com Bayer e Cribari–Neto (2015), a modelagem do parâmetro de dispersão (recíproco da precisão) pode ser de interesse na identificação de fontes de variabilidade no fenômeno em estudo. Além disso, Canterle *et al.* (2015) também concluem que a correta modelagem deste parâmetro por meio da seleção de covariáveis e funções de ligação têm influência direta na eficiência dos estimadores dos parâmetros de regressão da média. Estes fatos demonstram a real importância de se modelar corretamente esta estrutura de regressão. Porém, Souza T.C. *et al.* (2016) afirmam que modelar a variabilidade é um processo tipicamente mais complicado do que modelar a média da variável resposta, e como consequência disso, torna-se mais fácil se cometer erros de especificação nesta estrutura. Baseado nisso, Cribari–Neto e Souza (2012) propuseram estimadores do tipo sanduíche para o modelo de regressão beta, sendo esta uma alternativa para se realizar inferências precisas no submodelo da média sem necessariamente ter que se modelar a dispersão. As inferências obtidas utilizando esta abordagem são precisas mesmo sob dispersão variável. Uma outra alternativa para se evitar erros de especificação, caso seja importante para o estudo modelar a variabilidade, é a aplicação do teste *RESET* adaptado para a classe de modelos de regressão beta (LIMA, 2007; OLIVEIRA, 2013). Segundo os autores, este teste é útil na identificação de diversos tipos de erros de especificação que usualmente são cometidos.

4 Aplicação

Nesta seção apresentamos uma aplicação do modelo de regressão beta com dispersão variável aos dados da obesidade adulta nos Estados Unidos no ano de 2014. Estes dados são referentes ao estudo realizado por Souza S.A. *et al.* (2016). Aqui, nosso interesse consiste em comparar os resultados inferenciais obtidos através do uso de diferentes funções de ligação. Para esses dados temos que a variável resposta, *OB2014*, é a proporção de adultos obesos nos estados e totalizam 50 observações. As variáveis independentes utilizadas para explicar a obesidade nos estados foram: a porcentagem de residentes desempregados ou empregados em tempo parcial em 2014 (*DESEMP*), a porcentagem de adultos que consomem vegetais menos de uma vez por dia em 2011 (*VEGET*), a porcentagem de residentes que não tem cobertura

de seguro de saúde em 2014 (*DESCOB*), o escore de bem-estar em 2014 (*BST*), a porcentagem de fumantes de cigarro em 2012 (*FUM*) e a taxa de insegurança alimentar em 2013 (*INSEG*). Neste estudo, a especificação do modelo de regressão beta com dispersão variável pode ser definida da seguinte maneira:

$$g(\mu_t) = \beta_0 + \beta_1 DESEMP_t + \beta_2 VEGET_t + \beta_3 DESCOB_t + \beta_4 BST_t + \beta_5 FUM_t$$

$$h(\phi_t) = \gamma_0 + \gamma_1 INSEG_t + \gamma_2 DESCOB_t + \gamma_3 VEGET_t,$$

com $t = 1, \ldots, 50$. Além disso, temos que $g(\cdot)$ e $h(\cdot)$ são as funções de ligação utilizadas para modelar a média e precisão, respectivamente. Vale salientar que a seleção das covariáveis e funções de ligação se deu de modo que houvessem dois modelos de regressão concorrentes que pudessem ser comparados a partir de diversos métodos usualmente utilizados em estudos na área, sendo os modelos aqui analisados denominados de ajuste 1 e ajuste 2 (ver Tabela 5). Aplicamos o teste da razão de verossimilhanças (NEYMAN e PEARSON, 1928; SOUZA S.A. *et al.*, 2016) para testar a hipótese nula de que a dispersão dos dados é fixa versus a hipótese alternativa de que a mesma é variável, concluindo ao nível de significância de 5% a necessidade de uma estrutura para modelar a dispersão dos dados.

Tabela 5 - Estimativas dos parâmetros (Est.), erros-padrão (E.P.) e p-valores dos modelos considerando as funções de ligação loglog para a estrutura da média e log e sqrt para a estrutura da precisão

Ajuste	1			2		
Especificação	loglog(μ_t) e log(ϕ_t)			loglog(μ_t) e sqrt(ϕ_t)		
Parâmetro	Est.	E.P.	p-valor	Est.	E.P.	p-valor
β_0	0.546	0.267	0.041	0.853	0.446	0.056
β_1	-0.005	0.001	< 0.001	-0.006	0.003	0.012
β_2	0.010	0.002	< 0.001	0.009	0.002	< 0.001
β_3	0.004	0.001	0.004	0.005	0.002	< 0.001
β_4	-0.018	0.004	< 0.001	-0.023	0.006	< 0.001
β_5	0.008	0.002	< 0.001	0.008	0.002	< 0.001
γ_0	3.110	1.557	0.046	-17.409	18.094	0.336
γ_1	-0.370	0.094	< 0.001	–	–	–
γ_2	0.138	0.062	0.025	–	–	–
γ_3	0.332	0.059	< 0.001	2.146	0.835	0.010
λ	1349.416			11.509		

A Tabela 5 apresenta as estimativas, erros-padrão e p-valores obtidos após a modelagem dos dados considerando a função de ligação loglog para o submodelo da média, log para o submodelo da precisão no ajuste 1 e sqrt para o submodelo da precisão no ajuste 2. Nesta tabela são apresentados apenas os resultados inferenciais para aquelas variáveis cujo o parâmetro apresentou um p-valor do teste significativo. Desta forma, pôde-se verificar que para o ajuste 2, os parâmetros γ_0, γ_1 e γ_2 não foram significativos, sendo este modelo reajustado sem as covariáveis

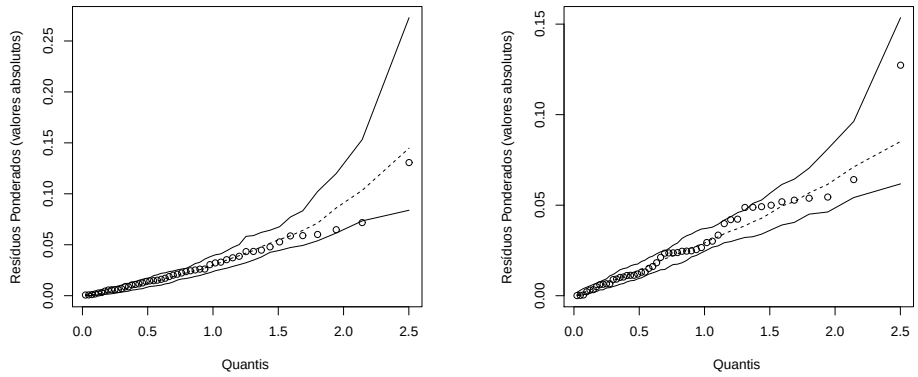
correspondentes aos parâmetros γ_1 e γ_2. Para cada modelo foi avaliado a razão entre os valores de máximo e mínimo da precisão (λ), que pode ser interpretado como uma medida de não-constância da precisão dos dados. Verificamos que a mudança na estrutura de regressão do parâmetro de precisão ocasionou uma grande variabilidade nesta medida, sendo que para o ajuste 1 o valor encontrado foi de 1349.416, enquanto que para o ajuste 2 este mesmo valor foi de 11.509, resultando em uma mudança considerável no grau de precisão. Além disso, avaliamos o teste *RESET* (LIMA, 2007; OLIVEIRA, 2013), o gráfico de probabilidade normal com envelopes simulados, o pseudo-R^2 e as medidas de influência para que fosse possível comparar as mudanças inferenciais ocorridas quando se utilizam diferentes formas de especificação na modelagem dos dados. Por fim, aplicamos o teste J (CRIBARI–NETO e LUCENA, 2015) para chegarmos ao modelo melhor especificado.

Ferrari e Cribari–Neto (2004) propuseram uma medida similar ao coeficiente de determinação para modelos de regressão beta, denominado pseudo-R^2. Esta medida permite avaliar a qualidade do ajuste dos modelos e apresenta valores contidos no intervalo $(0, 1)$, sendo que, quanto mais próximo de 1 melhor a capacidade explicativa do modelo proposto. Desta forma, para os ajustes 1 e 2 obtivemos respectivamente os valores de 0.7721 e 0.7834. Portanto, a mudança na estrutura de regressão do parâmetro de precisão ocasionou uma variação da capacidade explicativa dos modelos, sendo o ajuste 2 o que apresentou a maior capacidade explicativa.

Nós utilizamos o teste *RESET* adaptado para modelos de regressão beta (LIMA, 2007; OLIVEIRA, 2013) com o objetivo de testar a correta especificação dos modelos propostos. A hipótese nula sugere que o modelo testado está bem especificado versus a hipótese alternativa de que o mesmo está mal especificado. Para a realização do teste, consideramos o preditor linear estimado elevado a segunda potência ($\widehat{\eta}^2$) como variável de teste incluída no submodelo da média. Desta forma, obtivemos os p-valores de 0.1141 e 0.1522 referentes aos ajustes 1 e 2, respectivamente. Portanto, podemos concluir que tais modelos não apresentam erros na omissão de variáveis ou forma funcional incorreta aos níveis usuais de significância. Vale destacar aqui, que quando consideradas as funções de ligação cloglog e log para modelar as estruturas da média e da precisão, respectivamente, o teste *RESET* sugere a incorreta especificação do modelo ao nível de significância de 5% (p-valor = 0.0444), podendo o mesmo apresentar algum tipo de erro de especificação.

A Figura 2 apresenta os gráficos de probabilidade normal com envelopes simulados utilizando os resíduos ponderados (ESPINHEIRA *et al.*, 2008*a*). A partir destes gráficos é possível concluir que os ajustes avaliados apresentam os resíduos, em sua maioria, dentro das bandas de confiança dos envelopes simulados, mostrando a adequabilidade dos modelos independente da especificação utilizada. Valendo aqui ressaltar que os resíduos do ajuste 1 apresentaram menores desvios comparado aos do ajuste 2.

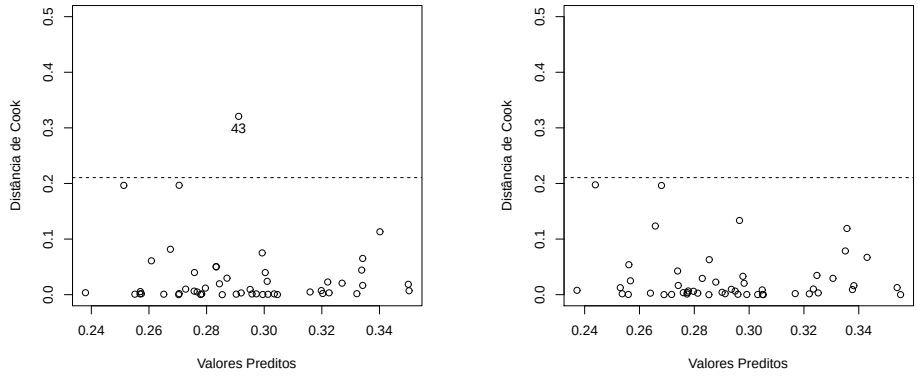
A partir da análise de influência para modelos de regressão beta (ESPINHEIRA *et al.*, 2008*a*) foi possível avaliar algumas medidas como distância



(a) loglog e log (b) loglog e sqrt

Figura 2 - Gráficos de probabilidade normal com envelopes simulados.

de Cook e alavancagem generalizada. A distância de Cook foi introduzida por Cook (1977) como uma medida capaz de quantificar o impacto de cada observação nas estimativas dos parâmetros. A Figura 3 apresenta os gráficos da distância de Cook versus os valores preditos, sendo possível visualizar que houveram diferenças na classificação das observações de acordo com a especificação utilizada. Para o ajuste 1 apenas a observação 43, referente ao estado do Texas, encontrou-se destacada

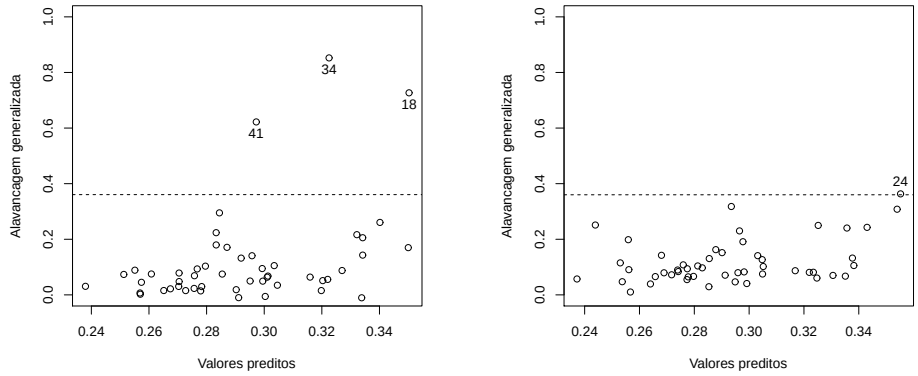


(a) loglog e log (b) loglog e sqrt

Figura 3 - Gráficos das distâncias de Cook.

frente as demais, sendo este estado o que apresentou uma das maiores porcentagens de residentes que não tinham cobertura de seguro de saúde (*DESCOB*). Por outro lado, para o ajuste 2 não foram identificamos pontos de influência.

A alavancagem generalizada foi proposta por Wei *et al.* (1998) como uma medida da importância individual de cada observação. A Figura 4 apresenta os gráficos da alavancagem generalizada versus os valores preditos, e a partir dela, é possível visualizar que a alteração na estrutura de regressão da precisão ocasionou mudanças nestes gráficos. Para o ajuste 1 temos que as observações 18, 34 e 41, referentes aos estados de Louisiana, North Dakota e South Dakota, respectivamente, foram consideradas pontos de alavanca, enquanto que para o ajuste 2 apenas a observação 24, referente ao estado do Mississipi, encontrou-se destacada frente as demais. Vale salientar que o estado de Louisiana apresentou a maior porcentagem de adultos que consumiam vegetais menos de uma vez ao dia (*VEGET*), enquanto que North Dakota apresentou a menor porcentagem de residentes desempregados ou empregados em tempo parcial e a menor taxa de insegurança alimentar (*INSEG*). O estado de South Dakota apresentou um dos maiores escores de bem-estar (*BST*), enquanto que o Mississipi apresentou a maior taxa de insegurança alimentar (*INSEG*) e uma das maiores porcentagens de adultos que consumiam vegetais menos de uma vez ao dia (*VEGET*).



(a) loglog e log (b) loglog e sqrt

Figura 4 - Gráficos de alavancagem generalizada.

Por fim, a Figura 5 apresenta o gráfico dos valores observados versus os valores estimados, permitindo assim a comparação entre as respostas médias estimadas obtidas a partir de cada modelo proposto. Como resultado, verificamos que as estimativas médias de *OB2014* se diferenciaram dependendo da especificação utilizada no modelo, contudo não se foi possível indicar um melhor modelo considerando esta medida, visto que as estimativas foram bem próximas dos valores reais em ambos os ajustes. Valendo ressaltar que o melhor ajuste seria aquele que

apresentasse os menores desvios, denotados por $\mu_t - \widehat{\mu}_t$, com $t = 1, \ldots, 50$, ou seja, apresentando observações mais próximas possível da reta.

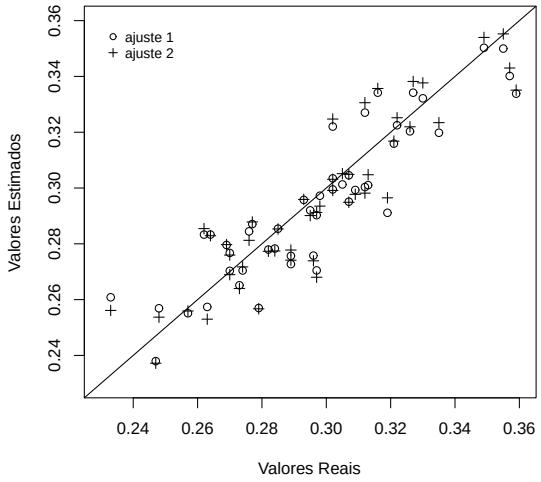


Figura 5 - Gráfico dos valores observados versus os valores estimados da variável obesidade adulta nos Estados Unidos em 2014, considerando os ajustes com diferentes funções de ligação.

Uma das maneiras de se decidir entre dois modelos de regressão beta não-encaixados qual que apresenta-se melhor especificado é utilizar o teste J adaptado para este tipo de modelo apresentado por Cribari–Neto e Lucena (2015). A aplicação deste teste é realizada de forma sequencial para cada modelo. Primeiramente, para testar a especificação do ajuste 1, incluímos no mesmo a estimativa do preditor linear do ajuste 2 como variável de teste, em seguida, testamos a significância do modelo por meio do teste da razão de verossimilhanças, obtendo assim um p-valor 0.0440. Em sequência, testamos a especificação do ajuste 2 incluindo a estimativa do preditor linear do ajuste 1 como variável de teste, obtendo um p-valor de 0.2350. Desta forma, concluímos aos níveis usuais de significância que o ajuste 2 apresentou uma melhor especificação, ou seja, a partir do teste J temos que o modelo melhor especificado é aquele com funções de ligação loglog e sqrt para modelar as estruturas de regressão da média e da precisão, respectivamente.

5 Conclusões

Neste artigo avaliamos o efeito de erros de especificação nas inferências do modelo de regressão beta com dispersão variável. Para isto, um estudo de simulação

foi realizado considerando diferentes cenários. Neste estudo, o modelo de regressão beta foi ajustado sob a especificação correta e incorreta. Em particular, seis tipos de erros de especificação foram avaliados, englobando tanto erros nos preditores quanto nas funções de ligação dos submodelos da média e da precisão, incluindo o caso em que a estrutura de regressão para o parâmetro de precisão é negligenciada erroneamente.

Verificamos através das taxas de rejeição que os modelos estimados com erro no preditor da precisão tenderam a apresentar valores mais distantes dos níveis nominais, comparados aos demais modelos. Além do mais, o modelo estimado com mais erros de especificação (erros nas funções de ligação e nos preditores) apresentou as maiores taxas de rejeição, considerando médias localizadas próximas a 0 e a 0.5. Em contrapartida, para valores de médias próximas a 1, errar nos preditores das duas estruturas de regressão (média e precisão) se mostrou um erro mais grave. Adicionalmente, negligenciar a estrutura de regressão do parâmetro de precisão, em amostras moderadas e grandes, se mostrou mais grave do que errar no preditor linear da média e nas funções de ligação dos dois submodelos, o que indica que para estes tipos de amostras é melhor permitir que o parâmetro de precisão varie ao longo das observações, enquanto que para amostras pequenas é menos grave que o mesmo permaneça constante. Em relação a análise dos resultados referentes as taxas de cobertura do parâmetro β_2, houve a confirmação de algumas das conclusões citadas acima. Por exemplo, nos modelos estimados com erro no preditor da precisão os resultados tenderam a ser mais distantes do valor esperado (95%) e o modelo estimado com erros nas funções de ligação e nos preditores dos dois submodelos de regressão, no geral, apresentou os piores resultados comparados aos demais. Em relação as medidas utilizadas para avaliar as estimativas para as respostas médias, o modelo estimado com precisão fixa apresentou, em alguns casos, os resultados mais distantes do valor esperado. Por fim, uma aplicação a dados reais foi realizada com o objetivo de verificar na prática os efeitos de diferentes formas de especificação no modelo de regressão beta com dispersão variável.

Observamos que, os erros de especificação que envolviam o preditor linear da estrutura de regressão do parâmetro de precisão tiveram uma influência considerável nas inferências do modelo de regressão beta. Este resultado confirma o que é encontrado na literatura. Uma das possíveis soluções para contornar este problema, caso seja de interesse no estudo a identificação de fontes de variabilidade, é o uso do teste *RESET* adaptado para a classe de modelos de regressão beta. Este teste é adequado para identificar possíveis erros de especificação que usualmente são cometidos. Uma outra solução, caso não se tenha tanto interesse em modelar a variabilidade, é o uso dos estimadores do tipo sanduíche propostos por Cribari–Neto e Souza (2012). Estes estimadores são adequados para os casos em que a estrutura de regressão que modela a variabilidade é negligenciada. Segundo os autores, as inferências obtidas através do uso destes estimadores apresentam bons resultados mesmo sob dispersão variável.

Agradecimentos

Agradecemos ao CNPq e à Capes pelo apoio financeiro e aos dois revisores e editores pelas sugestões.

OLIVEIRA, A.A.; SOUZA, T.C.; SOUZA, S.A. Misspecification in beta regression model with varying dispersion. *Rev. Bras. Biom.,* Lavras, v.36, n.3, p.601-624, 2018.

- *ABSTRACT: Our goal with this article is to evaluate the effects of misspecifications in the inferences of the beta regression model with varying dispersion. For this, a simulation study was carried out. In these simulations, the response variable was generated with beta distribution assuming known covariates and link functions, thus the model has been adjusted in the correct and incorrect specification, particularly considering six kinds of specification errors. We evaluate the effects of these errors through rejection rates and coverage rates in relation to one of the average submodel parameters and, in addition, we also evaluated the relative bias and the mean square error of the estimates mean responses. We verified from the results obtained, that the specification errors involving the linear predictor of the precision regression structure had a considerable influence on the model inferences. Finally, we performed an application to real data in order to compare the effects of different ways specification on the inferences of beta regression model with varying dispersion.*

- *KEYWORDS: Beta regression model; misspecification; rejection rates.*

Referências

ALMEIDA JUNIOR, P. M.; SOUZA, T. C. Estimativas de votos da presidente Dilma Roussef nas eleições presidenciais de 2010 sob o âmbito do Bolsa Família. *Ciência e Natura*, v.37, n.1, p.12-22, 2015.

ANDRADE, A. C. G. *Efeitos da especificação incorreta da função de ligação no modelo de regressão beta*, 2007. 88p. Dissertação (Mestrado em Ciências) - Universidade Federal de São Paulo, 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.; PALM, B. G.; BAYER, F. M. Efeitos da especificação incorreta das funções de ligação no modelo de regressão beta com dispersão variável. *Revista Brasileira de Biometria*, v.33, n.3, p.378-394, 2015.

COOK, R. D. Detection of influential observations in linear regression. *Technometrics*, v.19, n.1, p.15-18, 1977.

CRIBARI–NETO, F.; LUCENA, S. E. F. Nonnested hypothesis testing in the class of varying dispersion beta regression. *Journal of Applied Statistics*, v.42, p.967-985, 2015.

CRIBARI–NETO, F.; PEREIRA, T. L. Avaliação da eficiência de administrações municipais no estado de São Paulo: uma nova abordagem via modelos de regressão beta. *Revista Brasileira de Biometria*, v.31, n.2, p.270-294, 2013.

CRIBARI–NETO, F.; QUEIROZ, M. P. F. On testing inference in beta regressions. *Journal of Statistical Computation and Simulation*, v.84, p.183-203, 2014.

CRIBARI–NETO, F.; SOUZA, T. C. Testing inference in variable dispersion beta regressions. *Journal of Statistical Computation and Simulation*, v.82, 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.

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

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

ESPINHEIRA, P. L.; FERRARI, S. L. P.; CRIBARI–NETO, F. Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, v.52, p.4417-4431, 2008b.

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

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on $(0, 1)$: percentages, proportions and fractions. *Statistical Modelling*, v.3, p.193-213, 2003.

KLEIBER, C.; ZEILEIS, A. *Applied Econometrics with R*. New York: Springer, 2008. 222p.

LIMA, L. B. *Um teste de especificação correta para modelos de regressão beta*, 2007. 107p. Dissertação (Mestrado em Estatística) - Universidade Federal de Pernambuco, Recife, 2007.

LOOSE, L. H.; PALM, B. G.; BAYER, F. M. Avaliação dos estimadores do modelo de regressão beta com dispersão variável: um estudo de simulação. *Revista Eletrônica Matemática e Estatística em Foco*, v.2, n.1, p.14-24, 2014.

MCCULLAGH, P.; NEIDER, J. A. *Generalized Linear Models*, 2.ed. London: Chapman and Hall, 1989. 511p.

NEYMAN, J.; PEARSON, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, v.20, p.175-240, 1928.

OLIVEIRA, A. A.; SOUZA, T. C. Avaliação da proporção de crianças obesas beneficiadas pelo Programa Bolsa Família nas regiões do Brasil. *Revista Ciências Exatas e Naturais*, v.18, n.1, p.55-80, 2016.

OLIVEIRA, J. S. C. *Detectando má especificação em regressão beta*, 2013. 90p. Dissertação (Mestrado em Estatística) - Universidade Federal de Pernambuco, Recife, 2013.

PEREIRA, T. L.; CRIBARI–NETO, F. Detecting Model Misspecification in Inflated Beta Regressions. *Communications in Statistics – Simulation and Computation*, v.43, p.631-656, 2014.

PINTO, E. R.; PEREIRA, L. A.; RESENDE, L. O.; DESTRO FILHO, J. B. Modelos Estatísticos para estimação da área miocárdica sob risco de necrose. *Revista Brasileira de Biometria*, v.29, n.3, p.395-415, 2011.

PRESS, W.; TEUKOLSKY, S.; VETTERLING, W.; FLANNERY, B. *Numerical recipes in C: the art of scientific computing*. 2.ed. Cambridge University Press, 1992.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Fondation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2017.

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

SANT'ANNA, A. M. O.; CATEN, C. S. Modelagem da fração de não-conformes em processos industriais. *Pesquisa Operacional*, v.30, n.1, p.53-72, 2010.

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.; ROCHA, A. V. Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, v.54, p.348-366, 2010.

SOUZA, T. C.; CRIBARI-NETO, F. Intelligence, religiosity and homosexuality non-acceptance: Empirical evidence. *Intelligence*, v.52, p.63-70, 2015.

SOUZA, S. A.; OLIVEIRA, A. A.; SOUZA, T. C.; LIMA, C. M. B. L. Modelagem da proporção de obesos nos Estados Unidos utilizando modelo de regressão beta com dispersão variável. *Ciência e Natura*, v.38, n.3, p.1146-1156, 2016.

SOUZA, T. C.; PEREIRA, T. L.; CRIBARI-NETO, F.; LIMA, V. M. C. Testing inference in inflated beta regressions under model misspecification. *Communications in Statistics - Simulation and Computation*, v.45, p.625-642, 2016.

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

WALD, A. Test of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, v.54, n.3, p.426-482, 1943.

WEI, B.; HU, Y.; FUNG, W. Generalized leverage and its aplications. *Scandinavian Journal of Statistics*, v.25, n.1, p.25-37, 1998.

Received in 21.12.2016.

Approved after revised in 22.01.2018.