Artigo original *DOI: 10.5902/2179460X14203*

Ciência e Natura, Santa Maria, v. 37 n. 1, 2015, p. 1–11
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA 35 anos

Testes de especificação para a função de ligação em modelos lineares generalizados para dados binários

Test for link misspecification in generalized linear models for binary data

Diego Ramos Canterle*[1] e Fábio Mariano Bayer†[2]

[1]Curso de Bacharelado em Estatística, Universidade Federal de Santa Maria
[2]Departamento de Estatística e LACESM, Universidade Federal de Santa Maria

Resumo

Este trabalho aborda o problema de testar a correta especificação da função de ligação em modelos lineares generalizados para dados binários. Para realização do teste RESET de especificação, além de serem consideradas as tradicionais estatísticas da razão de verossimilhanças, de Wald e escore, propomos a utilização da recente estatística gradiente. A avaliação dos testes foi realizada por meio de simulações de Monte Carlo. Foram verificados os desempenhos em amostras de tamanho finito dos quatro testes considerados, em termos de tamanho e poder, assim como avaliadas as distribuições das estatísticas de teste em pequenas amostras. Pode-se verificar que os testes de especificação são influenciados pela função de ligação utilizada e pelo tamanho amostral considerado. O desempenho da estatística gradiente se mostrou superior, principalmente nos menores tamanhos amostrais. Uma aplicação a dados reais é apresentada com a finalidade de ilustração do teste proposto.

Palavras-chave: *estatística gradiente, função de ligação, modelos lineares generalizados, simulações de Monte Carlo, teste RESET.*

Abstract

This paper addresses the issue of check the correct specification of the link function in generalized linear models for binary data. To perform the RESET test we consider the likelihood ratio, Wald and score traditional statistics and we propose the use of the emerging gradient statistic. The performance evaluation of misspecification tests were performed using Monte Carlo simulations. The finite sample performance of the tests were evaluated in terms of size and power tests. It can be seen that the performance of the tests are influenced by the used link function and the sample size. The gradient statistic outperforms the traditional statistics, especially in smaller sample sizes. An empirical application to a real data set is considered for illustrative purposes.

Keywords: *generalized linear models, gradient statistic, link function, Monte Carlo simulations, RESET test.*

*D. R. Canterle: diegocanterle@gmail.com
†F. M. Bayer: bayer@ufsm.br
Recebido: 08/08/2014 Revisado: 05/09/2014 Aceito: 25/09/2015

1 Introdução

Os modelos lineares generalizados (MLG) (McCullagh e Nelder, 1989), introduzidos inicialmente por Nelder e Wedderburn (1972), são amplamente utilizados para modelar uma variável de interesse (y) que pertence à família exponencial. Nestes modelos, a média da variável resposta y é modelada por meio de uma estrutura de regressão que envolve parâmetros desconhecidos, covariáveis e uma função de ligação. Dentre os modelos de maior importância se encontram os modelos da família binomial, úteis para modelar variáveis binárias.

Os modelos de regressão para dados binários são utilizados quando a variável dependente do modelo assume apenas valores zero ou um. Esses modelos são amplamente utilizados em diversas áreas de aplicação, como pode ser visto em Harrell-Jr. et al. (1996); Wasserman e Pattison (1996); Chen e Randallb (1997); Ayalew e Yamagishi (2005). Um texto importante dedicado especificamente à esses modelos pode ser verificado no Capítulo 4 do livro de McCullagh e Nelder (1989).

A relação entre a média do componente aleatório, y, e o preditor linear do modelo é apresentada por meio de uma função de ligação. Para cada distribuição assumida para a variável dependente em um MLG existem diferentes funções de ligação que podem ser utilizadas (McCullagh e Nelder, 1989, Pag. 31). Neste trabalho, que aborda modelos binários, são consideradas as funções de ligação logit, probit, complemento log-log (cloglog) e Cauchy. Para seleção ou teste de adequacidade da função de ligação considerada existem diferentes alternativas. Usualmente a verificação da correta especificação da função ligação de um MLG é feita através de uma visualização gráfica de $\hat{z} = \hat{\eta} + \hat{H}(y - \hat{\mu})$ versus $\hat{\eta}$, em que $\hat{\eta}$ é o vetor de preditores lineares, $\hat{H}$ é a matriz de projeção ortogonal local, y é o vetor de valores observados da variável dependente e $\hat{\mu}$ é o vetor de valores estimados para a média (Demétrio, 2001; Paula, 2010). Alternativamente, como proposto por Hinkley (1985), é útil observar a redução no desvio após a inclusão da covariável $\hat{\eta}^2$ para verificar a especificação incorreta da função de ligação em um MLG. Outra alternativa seria utilizar o teste RESET (*Regression Specification Error Test*) de Ramsey (1969), e versões alternativas, os quais foram avaliados em Ramalho e Ramalho (2012) para modelos com respostas binárias, mas apenas em grandes amostras. Em outra classe de modelos, nos modelos de regressão beta, é possível verificar um maior número de estudos recentes a respeito do teste de especificação da função de ligação. Em Andrade (2007) é feita verificação do impacto da especificação incorreta da função de ligação no modelo de regressão beta através de estudos de simulação. Uma avaliação do desempenho do teste RESET para verificar a especificação correta no modelo de regressão beta é apresentada em Oliveira (2013) e para o modelo de regressão beta inflacionado em Pereira e Cribari-Neto (2013).

Para testar a correta especificação de um modelo de regressão, como a omissão ou excesso de covariáveis e/ou função de ligação incorreta, é usual a utilização do teste RESET (Ramsey, 1969). Este teste consiste em incluir $\hat{\eta}^2$ no modelo e testar se essa nova "variável" deve permanecer no modelo. Quando as inferências do modelo são feitas via máxima verossimilhança, como nos MLG's, é usual a utilização das seguintes estatísticas para testar a significâncias de covariáveis no modelo: razão de verossimilhanças (Neyman e Pearson, 1928), Wald (Wald, 1943) e escore (Rao, 1948). Uma referência clássica sobre estes testes é o artigo de Buse (1982). Outra possibilidade é o uso da recente estatística gradiente (Terrell, 2002). Todas as quatro estatísticas de teste citadas possuem distribuição nula assintótica qui-quadrado, em que valores críticos dessa distribuição de referência podem ser utilizados para realização dos testes. Contudo, em pequenas amostras os resultados inferenciais desses testes podem ser distorcidos, uma vez que as aproximações das distribuições nulas exatas das estatísticas de teste pela distribuição qui-quadrado podem ser pobres (Bayer e Cribari-Neto, 2013; Pereira e Cribari-Neto, 2013; Vargas et al., 2014). Neste sentido, torna-se importante avaliar numericamente o desempenho desses testes de hipóteses em amostras de tamanho finito.

Neste sentido, propomos a utilização da estatística gradiente para realização do teste RESET para testar a correta especificação de uma função de ligação em MLG para dados binários. A estatística gradiente é uma derivação das estatísticas de Wald e escore, porém de natureza distinta (Vargas et al., 2013). Muitos trabalhos sobre a estatística gradiente vem sendo desenvolvidos (Lemonte e Ferrari, 2012; Lemonte, 2013; Vargas et al., 2013). Uma grande vantagem da utilização da estatística gradiente é a facilidade de ser calculada, pois não utiliza a matriz de informação de Fisher, que em problemas complexos tem obtenção bastante custosa (Vargas et al., 2013), além de não exigir a inversão de matrizes. Neste trabalho serão avaliadas, por meio de simulação de Monte Carlo, o desempenho de quatro estatísticas de teste quando utilizadas para testar a exclusão da covariável $\hat{\eta}^2$ do modelo como proposto no teste RESET, verificando a correta especificação de modelos com resposta binária. Serão utilizadas as estatísticas da razão de verossimilhanças, Wald, escore e a recente estatística gradiente. A avaliação do desempenho dos testes se dará por meio das taxas de rejeição nula (tamanho) e das taxas de rejeição não-nula (poder) dos testes em amostras de tamanho finito. Também serão avaliadas as distribuições das estatísticas de teste em amostras

de tamanho finito e suas aproximações com a distribuição nula limite qui-quadrado. A melhor de nosso conhecimento, trabalhos que avaliam comparativamente o desempenho em pequenas amostras dos testes da razão de verossimilhanças, de Wald, escore e gradiente quando utilizados no teste RESET de especificação de modelos são inexistentes na literatura estatística.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta o modelo linear generalizado para dados binários, juntamente com uma breve descrição das funções de ligação consideradas. Na Seção 3 é introduzido o teste RESET de especificação, assim como as estatísticas de teste utilizadas. A Seção 4 apresenta os resultados numéricos. Uma aplicação à dados reais é introduzida na Seção 5. Finalmente, na Seção 6 são apresentadas as conclusões do trabalho.

2 O modelo

Algumas técnicas estatísticas podem ser englobadas em uma ampla classe de modelos de regressão, conhecidos como MLG. Esses modelos consideram uma variável resposta univariada y, k variáveis explicativas X_j, com $j = 1, \ldots, k$, uma amostra aleatória com n observações, um vetor k-dimensional de parâmetros desconhecidos β's e uma função de ligação $g(\cdot)$. Matricialmente, a estrutura de regressão pode ser escrita da seguinte forma:

$$g(\mu) = \eta = \mathbf{X}\beta, \tag{1}$$

em que $\mathbf{X}$ é a matriz de covariáveis com dimensão $n \times k$, β é o vetor k-dimensional de parâmetros desconhecidos a serem estimados, η é o vetor n-dimensional de preditores lineares e $\mu = (\mu_1, \ldots, \mu_n)^\top$ é o vetor de médias, tal que $\mathrm{E}(y_i) = \mu_i$, com $i = 1, \ldots, n$.

Em um MLG, o componente aleatório y_i é uma variável aleatória da família exponencial canônica com parâmetro de perturbação. A função densidade de probabilidade de y_i é dada por (Cordeiro e Demétrio, 2007):

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \tag{2}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas, θ_i é o parâmetro canônico e ϕ é o parâmetro de perturbação. O componente aleatório, com médias $\mu_1, \ldots, \mu_n$, se relaciona com o parâmetro canônico por meio de $\mathrm{E}(y_i) = \mu_i = b'(\theta_i)$.

O logaritmo da função de verossimilhança é dado por:

$$\ell(\beta; y_i) = \phi^{-1} \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi). \tag{3}$$

Tomando a primeira derivada da função de log-verossimilhança dada em (3) em relação a cada um dos parâmetros β_j do modelo em (1), temos o vetor escore:

$$\mathbf{U} = \phi^{-1}\mathbf{X}^T\mathbf{W}\mathbf{G}(y - \mu),$$

em que $\mathbf{G} = \mathrm{diag}\{d\eta_1/d\mu_1, \ldots, d\eta_n/d\mu_n\} = \mathrm{diag}\{g'(\mu_1), \ldots, g'(\mu_n)\}$ e $\mathbf{W} = \mathrm{diag}\{w_1, \ldots, w_n\}$ uma matriz diagonal de pesos, sendo $w_i = \dfrac{1}{V_i}\dfrac{d\mu_i}{d\eta_i}$ com $V_i = \dfrac{d\mu_i}{d\theta_i}$ (Cordeiro e Demétrio, 2007, Pag. 38).

A matriz de informação de Fisher, importante para inferência intervalar e testes de hipóteses em grandes amostras, é dada por:

$$\mathbf{K} = \phi^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}.$$

A função de ligação $g(\cdot)$ relaciona a média μ_i ao preditor linear η_i, isto é, $\eta_i = g(\mu_i)$, sendo $g(\cdot)$ uma função monótona e duas vezes diferenciável (Koenker e Yoon, 2009; Bayer e Cribari-Neto, 2013). Para uma determinada distribuição, se a função de ligação é tal que $\theta_i = \eta_i$, então está garantida a existência de uma estatística suficiente de dimensão igual a dimensão de β. Nestes casos, esta função é chamada de função de ligação canônica. O uso de funções de ligação canônicas tem as vantagens de ter uma escala adequada para a modelagem com interpretação prática dos parâmetros do modelo e de simplificar os algoritmos de estimação (McCullagh e Nelder, 1989).

2.1 O modelo binomial

Para modelos com variável dependente binária é comum assumir distribuição binomial com apenas um evento para y_i. Tal distribuição tem a seguinte função de probabilidade:

$$f(y_i; \mu_i) = \mu_i^{y_i}(1 - \mu_i)^{1-y_i}. \tag{4}$$

Essa função pode ser colocada na forma geral da família exponencial dada em (2), considerando $\phi = 1$, $\theta_i = \log\left(\dfrac{\mu_i}{1-\mu_i}\right)$, $b(\theta_i) = \log(1 + \exp(\theta_i))$, $\mu_i = \dfrac{\exp(\theta_i)}{1+\exp(\theta_i)}$ e $c(y_i, \phi) = 0$.

Nota-se que no modelo binomial a função de ligação canônica é a função logit, dada por:

$$g_1(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right).$$

Porém, existem outras funções de ligação além da canônica que associam os valores das médias μ_i, no intervalo padrão $(0,1)$, aos valores do preditor linear η_i, na reta real. Algumas delas são a probit, cloglog e Cauchy, dadas, respectivamente, por (Koenker e Yoon, 2009):

$$g_2(\mu_i) = \Phi^{-1}(\mu_i),$$

$$g_3(\mu_i) = \log[-\log(1-\mu_i)],$$
$$g_4(\mu_i) = \tan[\pi(\mu_i - 0{,}5)],$$

em que Φ é a função distribuição acumulada da distribuição normal padrão.

A seleção da função de ligação que será utilizada em uma modelagem com MLG's merece atenção especial. A utilização de uma função de ligação inadequada pode levar a erro do cálculo do componente de desvio (*deviance*) e também ter inferências incorretas sobre os β's, acarretando uma interpretação errada do modelo (McCullagh e Nelder, 1989, Pag. 401). Desta forma, o pesquisador terá uma falsa interpretação da realidade, causando grandes problemas decorrentes da tomada de decisão sobre o assunto em estudo.

A Figura 1 apresenta uma comparação gráfica das funções de ligação consideradas. Podemos verificar que a função de ligação cloglog é assimétrica, diferentemente das outras três funções. Para $\eta \approx 0$, as funções de ligação logit, probit e Cauchy levam a valores muito próximos de μ. Considerando $\eta \approx -1{,}5$ as funções de ligação logit, cloglog e Cauchy são muito semelhantes. Nota-se também que todas as funções são praticamente indistinguíveis quando valores de μ são muito próximos dos extremos do intervalo padrão $(0{,}1)$.

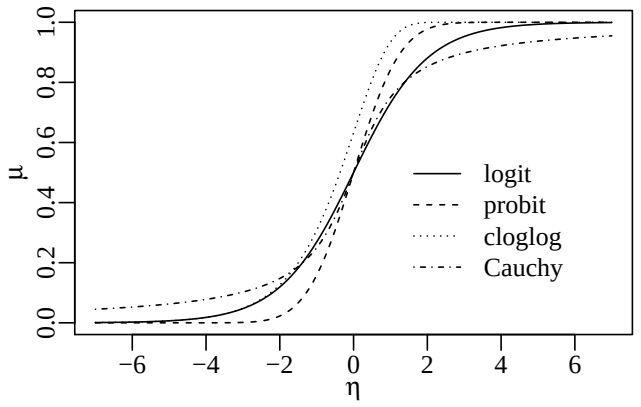


Figura 1: Gráfico comparativo entre as funções de ligação.

3 Teste RESET

O teste RESET (Ramsey, 1969) é realizado em duas etapas. A primeira consiste em incluir os valores preditos pelo modelo ajustado ao quadrado ($\widehat{\eta}^2$) como covariável, da seguinte forma:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \beta_{k+1}\widehat{\eta}_i^2,$$

em que k é o número de covariáveis do modelo. A segunda etapa consiste em testar as seguintes hipóteses:

$$\mathcal{H}_0 : \beta_{k+1} = 0 \text{ (corretamente especificado)}$$
$$\mathcal{H}_1 : \beta_{k+1} \neq 0 \text{ (incorretamente especificado).}$$

No caso do teste da correta especificação da função de ligação, considera-se que se a hipótese nula não é rejeitada, então a função de ligação está corretamente especificada. Caso contrário, sob rejeição da hipótese nula, conclui-se que a função de ligação está incorretamente especificada e deve ser selecionada outra função de ligação para os dados em estudo.

3.1 Estatísticas de teste

Para realização do teste RESET descrito na seção anterior, serão consideradas quatro estatísticas de teste, que são: a da razão de verossimilhanças (RV) (Neyman e Pearson, 1928), a de Wald (W) (Wald, 1943), a escore (S) (Rao, 1948) e a gradiente (G) (Terrell, 2002). Essas estatísticas são dadas, respectivamente, por:

$$\mathrm{RV} = 2[\ell(\widehat{\beta}) - \ell(\widetilde{\beta})],$$
$$\mathrm{W} = (\widehat{\beta} - \widetilde{\beta})\widehat{\mathbf{K}}_\beta(\widehat{\beta} - \widetilde{\beta}),$$
$$\mathrm{S} = \widetilde{\mathbf{U}}_\beta^\top \widetilde{\mathbf{K}}_\beta^{-1}\widetilde{\mathbf{U}}_\beta,$$
$$\mathrm{G} = \widetilde{\mathbf{U}}_\beta^\top(\widehat{\beta} - \widetilde{\beta}),$$

em que $\ell(\beta)$ é a log-verossimilhança, $\widehat{\beta}$ é o estimador de máxima verossimilhança dos parâmetros que indexam o modelo irrestrito, $\widetilde{\beta}$ é o estimador de máxima verossimilhança dos parâmetros que indexam o modelo restrito (sob $\mathcal{H}_0$), $\widetilde{\mathbf{K}}_\beta$ é a matriz informação de Fisher avaliada em $\widetilde{\beta}$, $\widehat{\mathbf{K}}_\beta$ é a matriz informação de Fisher sob hipótese alternativa e $\widetilde{\mathbf{U}}_\beta$ é o vetor escore do modelo restrito.

Sob condições usuais de regularidade e sob hipótese nula, as estatísticas de teste possuem distribuição assintótica χ^2_r, sendo r o número de restrições impostas por $\mathcal{H}_0$ (Vargas et al., 2014). No caso do teste RESET considerado, temos $r = 1$. Se o valor da estatística de teste for menor que o quantil usual da χ^2_r, não rejeita-se $\mathcal{H}_0$ e assume-se que o modelo está corretamente especificado. Caso contrário, rejeita-se $\mathcal{H}_0$ e assume-se que a função de ligação não é adequada e o teste deve ser realizado novamente utilizando outra função de ligação.

4 Avaliação numérica

Neste experimento foram avaliados o tamanho e o poder das implementações das estatísticas da razão de verossimilhanças, Wald, escore e gradiente do teste RESET. Foram utilizadas 50.000 réplicas de Monte Carlo e considerados tamanhos amostrais iguais a $n = 60$, $n = 100$,

Tabela 1: Taxas de rejeição nula (percentuais) do teste RESET, com dados gerados apartir da Equação 5.

	$\alpha = 10\%$					$\alpha = 5\%$					$\alpha = 1\%$				
Estat \ n	60	100	200	500	1000	60	100	200	500	1000	60	100	200	500	1000
	logit														
W	6,06	7,84	8,60	9,43	9,95	2,39	3,44	4,03	4,49	4,89	0,28	0,53	0,67	0,76	0,94
RV	8,10	9,12	9,40	9,81	10,23	3,88	4,50	4,65	4,82	5,09	0,81	0,81	0,90	0,93	1,03
S	6,82	8,65	9,12	9,64	10,05	3,14	4,15	4,52	4,72	5,00	0,77	0,85	0,88	0,88	0,98
G	9,05	9,57	9,68	10,01	10,33	4,77	4,95	4,93	4,96	5,19	1,22	1,07	1,05	1,02	1,07
	probit														
W	6,63	6,86	7,99	8,56	9,77	2,39	3,24	3,60	3,98	4,72	0,33	0,55	0,64	0,73	0,81
RV	8,34	7,42	8,59	9,10	10,34	3,85	3,44	3,97	4,46	5,24	0,71	0,61	0,66	0,80	1,09
S	7,05	7,25	8,39	8,79	9,79	3,26	4,24	4,24	4,28	4,82	1,52	1,51	1,13	0,99	0,93
G	9,63	8,41	9,15	9,52	10,63	5,23	4,24	4,49	4,82	5,56	0,95	0,96	0,90	1,01	1,28
	cloglog														
W	8,52	9,58	9,54	9,97	10,04	3,88	4,62	4,60	4,84	4,93	0,72	0,83	0,89	1,00	0,98
RV	10,42	10,24	10,01	10,28	10,23	5,18	5,12	5,13	5,10	5,04	1,10	0,96	1,03	1,04	1,02
S	9,08	9,82	9,73	10,03	10,11	4,24	4,76	4,81	4,93	4,95	0,86	0,85	0,91	0,99	0,97
G	11,36	10,45	10,24	10,42	10,23	6,16	5,38	5,35	5,25	5,12	1,73	1,10	1,20	1,14	1,09
	Cauchy														
W	5,09	5,91	6,16	8,23	9,60	1,16	2,50	2,17	3,44	4,75	0,00	0,18	0,31	0,55	0,99
RV	8,69	8,95	9,35	9,68	10,10	4,10	4,36	4,64	4,76	5,11	0,71	0,78	0,89	0,94	1,05
S	7,33	7,95	8,49	9,31	9,96	3,15	3,60	3,83	4,21	4,98	0,42	0,48	0,52	0,72	1,02
G	9,09	9,56	10,48	10,31	10,30	5,09	5,20	5,79	5,44	5,28	1,65	1,61	1,76	1,32	1,18

Tabela 2: Taxas de rejeição nula (percentuais) do teste RESET, com dados gerados apartir da Equação 6.

	$\alpha = 10\%$					$\alpha = 5\%$					$\alpha = 1\%$				
Estat \ n	60	100	200	500	1000	60	100	200	500	1000	60	100	200	500	1000
	logit														
W	9,14	9,47	9,59	10,03	10,03	3,93	4,31	4,64	5,05	4,95	0,43	0,61	0,74	0,96	0,97
RV	10,99	10,53	10,21	10,24	10,14	5,76	5,33	5,24	5,22	5,06	1,86	1,12	1,05	1,09	1,01
S	10,23	10,18	9,96	10,16	10,10	5,01	4,98	5,01	5,17	5,01	0,86	0,92	0,93	1,06	1,00
G	11,65	10,95	10,39	10,30	10,18	6,41	5,69	5,43	5,32	5,09	1,62	1,33	1,18	1,14	1,03
	probit														
W	9,91	10,19	10,23	10,17	10,12	4,41	4,83	4,99	5,14	4,96	0,55	0,77	0,91	0,99	0,98
RV	11,34	11,22	10,59	10,30	10,19	5,92	5,78	5,39	5,04	5,26	1,34	1,26	1,10	1,10	0,98
S	10,30	10,38	10,34	10,18	10,11	4,96	5,01	5,12	5,15	4,94	0,96	0,94	0,97	1,07	1,00
G	12,87	11,73	10,83	10,41	10,28	6,66	6,28	5,60	5,35	5,09	1,87	1,60	1,25	1,13	1,03
	cloglog														
W	9,37	8,77	10,02	10,19	10,02	3,73	3,77	4,78	4,98	5,06	0,28	0,48	0,74	0,93	0,92
RV	10,70	10,46	10,40	10,52	10,18	5,50	5,17	5,29	5,24	5,28	1,06	1,06	1,05	1,08	1,02
S	7,48	7,53	9,38	9,96	9,85	3,34	3,51	4,54	4,93	4,98	0,67	0,77	0,93	1,11	0,92
G	12,23	11,76	10,76	10,77	10,33	6,87	6,36	5,66	5,50	5,39	1,82	1,69	1,30	1,21	1,07
	Cauchy														
W	6,07	7,72	8,52	9,86	9,71	1,40	2,78	3,62	4,74	4,83	0,00	0,11	0,44	0,76	0,95
RV	10,87	10,50	10,11	10,39	9,82	5,66	5,39	5,01	5,26	4,97	1,98	1,10	1,02	1,01	1,02
S	10,21	10,14	9,83	10,30	9,82	5,12	4,91	4,92	5,19	4,91	0,89	0,92	0,94	0,96	0,99
G	11,30	10,79	10,34	10,46	9,88	6,19	5,60	5,22	5,31	5,01	1,65	1,42	1,19	1,06	1,02

Tabela 3: Taxas de rejeição nao-nula (percentuais) do teste RESET, com dados gerados apartir da Equação 5.

Estat \ n	60	100	200	500	1000	60	100	200	500	1000	60	100	200	500	1000
	gerada com logit														
	probit					clogclog					Cauchy				
W	2,65	3,76	4,36	4,91	5,23	3,11	3,74	4,25	4,53	4,98	0,66	1,99	2,28	4,83	9,23
RV	3,88	4,60	4,80	5,05	5,23	4,12	4,58	4,77	4,93	5,25	4,77	5,25	6,42	9,53	13,72
S	3,03	4,33	4,77	5,13	5,38	3,16	4,05	4,33	4,68	5,08	3,47	4,09	5,18	8,52	12,88
G	4,81	5,06	5,03	5,13	5,30	5,13	5,15	5,20	5,16	5,42	6,18	6,76	8,25	11,58	15,56
	gerada com probit														
	logit					clogclog					Cauchy				
W	1,75	2,45	2,65	3,56	5,67	2,02	2,61	2,62	3,86	6,83	0,10	0,63	1,56	17,67	49,55
RV	3,68	3,03	3,75	5,00	7,40	3,98	3,23	4,13	5,73	9,11	5,18	6,38	14,60	36,16	66,79
S	2,92	3,25	3,11	3,84	6,00	2,83	2,89	2,98	4,21	10,27	2,99	3,46	11,41	33,93	66,29
G	4,95	4,04	4,75	5,77	8,29	5,50	4,45	5,16	6,87	10,27	7,10	10,51	19,48	41,51	70,92
	gerada com cloglog														
	logit					probit					Cauchy				
W	3,45	4,36	4,64	5,08	5,32	3,88	4,74	5,03	5,67	6,34	1,42	2,86	2,98	4,99	9,99
RV	5,12	5,03	5,13	5,17	5,31	5,32	5,27	5,29	5,71	6,25	5,33	5,50	6,67	9,00	13,79
S	4,46	4,92	4,97	5,30	5,40	4,62	5,16	5,35	6,03	6,52	3,90	4,73	5,70	8,18	13,28
G	6,04	5,35	5,33	5,25	5,28	6,37	5,64	5,48	5,75	6,23	6,62	6,07	8,28	10,73	15,52
	gerada com Cauchy														
	logit					probit					clogclog				
W	3,24	3,94	4,46	5,86	7,07	3,56	4,26	4,79	6,36	7,62	3,94	4,46	4,58	5,65	6,74
RV	4,60	4,74	4,73	5,72	6,78	4,71	4,93	5,02	6,31	7,51	4,57	4,79	4,62	5,29	6,29
S	4,35	4,70	4,92	6,03	7,16	4,34	4,94	5,29	6,77	7,88	4,18	4,51	4,63	5,56	6,63
G	5,24	5,09	4,86	5,56	6,66	5,50	5,31	5,13	6,25	7,46	5,12	4,99	4,80	5,08	6,08

$n = 200$, $n = 500$ e $n = 1000$. Em cada réplica de Monte Carlo gerou-se n ocorrências da variável aleatória y_i, com $i = 1, \ldots, n$, com função de probabilidade dada em (4) e parâmetro de média definida por $\mu_i = g^{-1}(\eta_i)$, em que:

$$\eta_i = -1{,}5 + x_{i1} + (-1)x_{i2}, \quad (5)$$

$$\eta_i = 0{,}5 + x_{i1} + (-1)x_{i2}. \quad (6)$$

As covariáveis x_1 e x_2 foram geradas da distribuição uniforme (0,1) e consideradas constantes durante todas as réplicas. Para $g(\cdot)$ foram consideradas as funções de ligação logit, probit, cloglog e Cauchy. Todas as implementações computacionais foram realizadas utilizando a linguagem R (R Development Core Team, 2012). Para os ajustes dos modelos foi utilizada a função `glm()`.

A Tabela 1 apresenta os resultados de tamanho dos testes com os dados gerados a partir da Equação (5). Os resultados evidenciam, em geral, uma menor distorção de tamanho quando considerada a estatística gradiente, especialmente em amostras menores. A estatística RV também mostra bom desempenho, mas principalmente nas amostras maiores.

A partir da Tabela 2 podemos visualizar os resultados do tamanho dos testes com os dados gerados a partir da Equação (6). Os resultados, em geral, apresentam as menores distorções quando é considerada a estatística S. Em pequenas amostras observa-se que a estatística gradiente não possui consideráveis distorções de tamanho. Como esperado, todos os testes apresentam melhores resultados com o aumento do tamanho da amostra em todos os cenários.

Como resultados gerais das simulações de tamanho dos testes, nota-se que em praticamente todos os resultados da Tabela 1 a estatística gradiente apresenta menor distorção, enquanto que as demais estatísticas apresentam pobre desempenho em alguns casos. Por outro lado, os resultados da Tabela 2 não evidenciam o melhor desempenho da estatística gradiente, mas, mesmo nesses casos, sua distorção de tamanho não é considerável. Ou seja, de forma geral, a estatística gradiente mostra-se adequada, mesmo quando não é a melhor. Essa característica não é verificada para as outras estatísticas de teste.

As Tabelas 3 e 4 apresentam os resultados de poder dos testes, considerando $\alpha = 0{,}05$, quando negligenciada a correta especificação da função de ligação. Esses resultados mostram desempenhos pobres dos testes em alguns cenários. Verificam-se resultados razoáveis para os dois cenários a seguir: (i) para dados gerados a partir da Equação (5) com a função de ligação probit, e testados com a função de ligação Cauchy; (ii) para dados gerados

Tabela 4: Taxas de rejeição nao-nula (percentuais) do teste RESET, com dados gerados apartir da Equação 6.

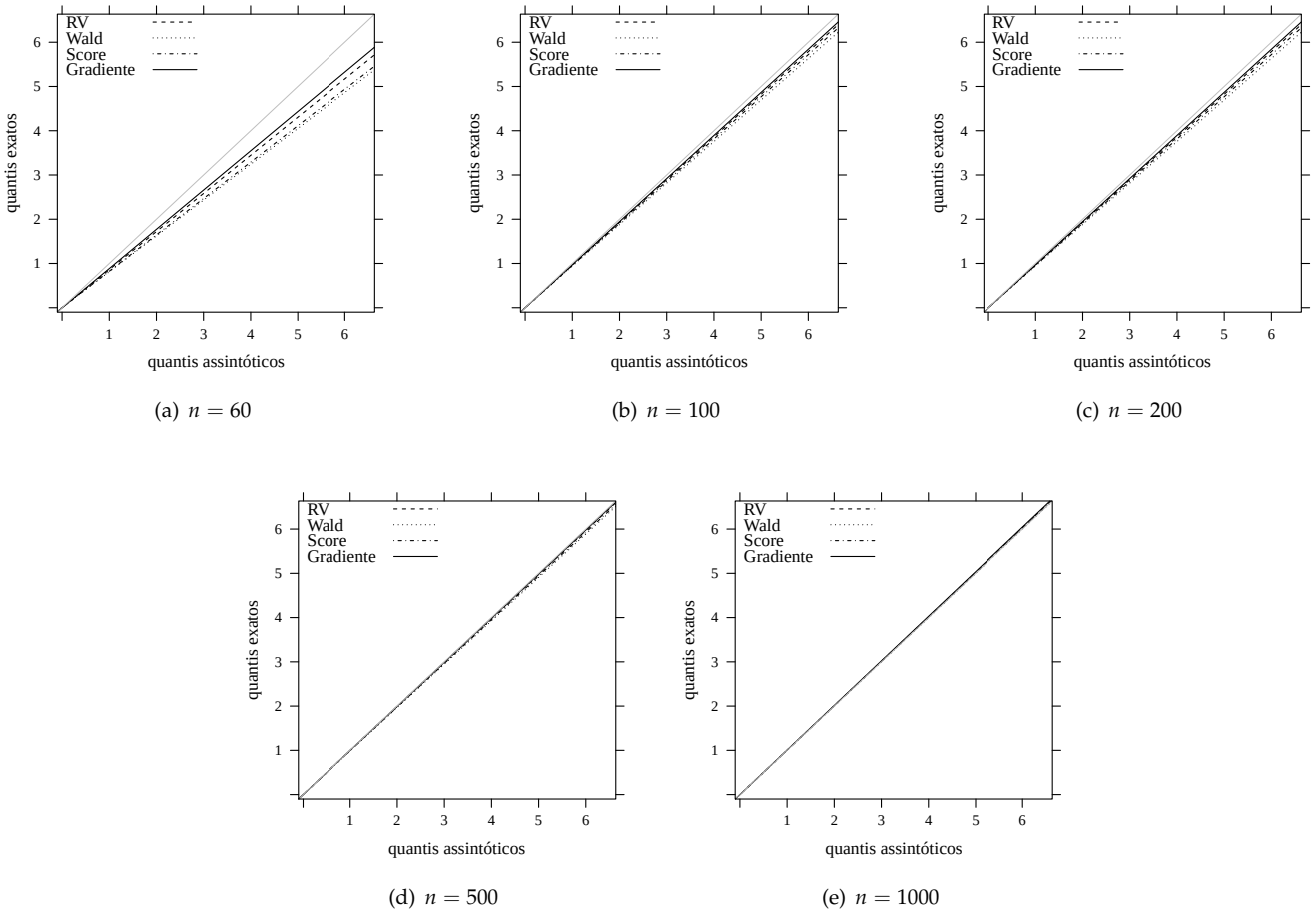
Estat \ n	60	100	200	500	1000	60	100	200	500	1000	60	100	200	500	1000
	gerada com logit														
	probit					clogclog					Cauchy				
W	4,54	4,79	4,96	5,22	5,07	4,54	4,91	5,01	5,70	6,32	1,69	2,69	3,59	4,91	5,96
RV	5,86	5,45	5,27	5,30	5,13	5,99	5,57	5,45	6,05	6,59	5,72	5,37	5,26	5,68	6,98
S	4,95	5,00	5,06	5,25	5,06	4,95	5,19	5,29	6,10	6,69	4,92	5,03	5,10	5,59	6,95
G	6,46	5,69	5,42	5,34	5,17	6,67	5,93	5,71	6,20	6,72	6,52	6,04	5,70	5,59	7,39
	gerada com probit														
	logit					clogclog					Cauchy				
W	3,87	4,26	4,80	5,11	5,41	4,51	5,25	5,61	7,26	10,63	1,64	2,77	5,03	11,20	24,25
RV	5,72	5,60	5,44	5,52	5,80	6,07	6,21	6,16	7,66	11,05	5,92	6,74	8,42	14,75	29,07
S	4,91	5,00	5,13	5,26	5,50	5,10	5,72	6,25	8,12	11,65	5,18	6,21	8,32	15,08	29,82
G	6,44	6,19	5,72	5,74	5,94	7,01	6,71	6,48	7,92	11,24	6,03	7,89	9,05	15,62	30,39
	gerada com cloglog														
	logit					probit					Cauchy				
W	2,42	3,02	5,91	10,65	20,17	3,16	3,40	5,34	7,43	11,07	0,51	1,26	9,87	38,61	78,57
RV	5,80	6,54	8,09	12,81	22,56	5,81	5,95	6,71	8,54	12,09	7,63	12,24	21,02	51,44	85,34
S	3,95	3,89	6,56	10,86	20,42	3,70	3,39	5,19	6,90	10,61	5,73	9,99	20,55	52,45	86,73
G	7,21	8,38	9,09	13,84	23,65	7,27	7,58	7,49	9,12	12,67	9,27	16,33	23,78	54,55	86,60
	gerada com Cauchy														
	logit					probit					clogclog				
W	4,18	4,67	5,34	6,70	8,88	4,95	5,27	5,75	7,29	10,25	4,97	5,43	6,40	9,38	17,09
RV	5,97	5,64	5,75	6,77	8,75	6,16	5,79	6,01	7,36	10,22	6,48	6,35	7,09	10,10	17,85
S	5,41	5,33	5,65	6,86	8,96	5,45	5,46	5,90	7,43	10,53	5,66	6,04	7,05	10,40	17,86
G	6,54	5,97	5,88	6,77	8,67	6,69	6,04	6,08	7,35	10,19	7,22	6,82	7,49	10,42	18,24

a partir da Equação (6) com a função de ligação cloglog, e testados com a função de ligação Cauchy. Em (i), considerando a estatística gradiente e tamanhos amostrais iguais a 500 e a 1000, por exemplo, os testes apresentam poderes de 41,51% e 70,92%, respectivamente. Já no cenário (ii), e nos mesmos casos anteriores, os poderes são de 54,55% e 86,60%. O melhor desempenho nesses cenários pode ser justificado pelo fato das funções de ligação terem maiores diferenças nos valores de μ para os mesmos valores de η (conf. Figura 1).

Quando o poder do teste não aumenta consideravelmente com o aumento do tamanho amostral, ou quando os poderes são muito baixos, isso sinaliza que as funções de ligação são muito próximas em certos valores de η em função de μ. Nestes casos, apesar dos testes terem baixa probabilidade de rejeição da hipótese incorreta de que o modelo está corretamente especificado, pelas semelhanças dessas funções nestes valores, as conclusões em uma modelagem empírica não são comprometidas. A Figura 1 mostra que para valores de η próximos de zero as funções de ligação logit, probit e Cauchy levam a valores de μ muito próximos. Neste caso, a relação entre os parâmetros β da estrutura de regressão e μ também não é comprometida.

Vale destacar que a implementação Gradiente do teste RESET tem poder considerável para tamanhos amostrais a partir de 500. Nenhuma das quatro estatísticas de teste avaliadas nesse trabalho mostram-se muito úteis para testar correta especificação da função de ligação através do teste RESET para tamanhos amostrais menores que 500. Os resultados evidenciando pouco poder em pequenas amostras são contribuições importantes deste estudo, uma vez que diversas referências clássicas sugerem uso do teste RESET para testar a correta especificação da função de ligação em MLG, mas poucos avaliam seu real desempenho.

Para análise das distribuições das estatísticas de teste em pequenas amostras, apresentam-se a Tabela 5 e a Figura 2. A Tabela 5 mostra os valores da média ($\overline{x}$), variância (s^2), assimetria (A), curtose (C) e quantis 0,90 ($Q90$), 0,95 ($Q95$) e 0,99 ($Q99$) das estatísticas de teste, assim como os valores de referência da distribuição nula limite qui-quadrado. Já a Figura 2 apresenta os gráficos quantil-quantil (QQ-plot) entre os quantis exatos (empíricos) das distribuições das estatísticas e os quantis assintóticas qui-quadrado. Nestes resultados é desejável que a distribuição nula limite qui-quadrado forneça boas aproximações para as distribuições exatas das estatísticas de teste. A estatística gradiente apresenta os valores de média, variância e quantis mais próximos dos valores



(a) $n = 60$ (b) $n = 100$ (c) $n = 200$

(d) $n = 500$ (e) $n = 1000$

Figura 2: Gráficos quantil-quantil das estatísticas de teste, considerando dados gerados a partir da Equação 5, com função de ligação logit e diferentes tamanhos amostrais n.

da distribuição nula limite. Para assimetria e curtose a estatística da razão de verossimilhanças tem melhor aproximação. Os gráficos Q-Q plot, da Figura 2, evidenciam uma melhor aproximação da distribuição nula limite à distribuição exata da estatística G. Essa verificação da melhor aproximação da distribuição da estatística G pela qui-quadrado justifica o melhor desempenho da mesma em termos de tamanho do teste.

Tabela 5: Medidas descritivas e quantis das estatísticas de teste, considerando dados gerados a partir da Equação 5, com função de ligação logit e $n = 100$.

	$\overline{x}$	s^2	A	C	$Q90$	$Q95$	$Q99$
χ^2_1	1,000	2,000	2,828	15,000	2,705	3,841	6,635
W	0,894	1,479	2,667	13,346	2,391	3,339	5,686
RV	0,954	1,811	2,815	14,579	2,563	3,671	6,270
S	0,941	1,825	3,142	18,605	2,507	3,564	6,365
G	0,986	2,082	3,178	19,378	2,635	3,824	6,784

Em geral, nota-se um bom comportamento da estatística G para realização do teste RESET, mostrando resultados de taxa de rejeição nula equivalentes ou superiores comparativamente às demais estatísticas de teste. Ademais, o teste gradiente mostrou-se o mais poderoso nos cenários considerados. Dessa forma, sugere-se a utilização dessa estatística para testar a correta especificação da função de ligação em modelos para dados binários. A próxima seção apresenta um exemplo de utilização do teste gradiente em dados reais.

5 Aplicação

Nesta seção é apresentada uma aplicação à dados reais do teste gradiente. Os dados considerados referem-se a processos infecciosos pulmonares em 175 pacientes do Setor de Anatomia e Patologia do Hospital Heliópolis em São Paulo. O período considerado foi de 1970 a 1982. Todos esses dados estão disponibilizados e descritos em Paula (2010). A variável de interesse, y, é uma variá-

Tabela 6: Modelo ajustado para os dados de processo infeccioso pulmonar considerando função de ligação logit.

Variável	Estimativa	Erro-padrão	p-valor
Intercepto	-3,74	0,93	<0,001
x_1	0,85	0,44	0,054
x_2	0,06	0,01	<0,001
x_3	-1,79	0,39	<0,001

vel binária que recebe valor 1 quando há presença de processo infeccioso pulmonar e zero caso contrário. As covariáveis considerados são: sexo (x_1), que recebe valor 1 se o paciente é do sexo feminino e 0 se é masculino, idade do paciente em anos (x_2), intensidade da célula histiócitos-linfócitos (x_3), que assume valor 0 quando há baixa intensidade e 1 quando há alta intensidade e intensidade da célula fibrose-frouxa (x_4), que é igual a 0 quando há baixa intensidade e igual a 1 para alta intensidade.

Para ajuste do modelo considerou-se a função de ligação canônica (logit) e um modelo com todas as covariáveis candidatas. As variáveis foram selecionadas por meio de um algoritmo *stepwise*, utilizando o critério de informação de Akaike (AIC) (Akaike, 1974). O modelo binário selecionado foi o seguinte:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

em que $i = 1,\ldots,175$.

No modelo ajustado foi considerada a estatística gradiente para realização do teste RESET, testando a correta especificação da função de ligação. Nesse teste obteve-se p-valor=0,9993. Portanto, não rejeita-se a hipótese nula e conclui-se que a função de ligação do modelo está corretamente especificada. A Tabela 6 apresenta o modelo ajustado. Nota-se que as covariáveis x_1 (sexo) e x_2 (idade) influenciam positivamente para presença de um processo infeccioso pulmonar, enquanto a covariável x_3 influencia negativamente. A influência negativa de x_3 na variável dependente está de acordo com o que é discutido em Paula (2010), uma vez que a chance de processo maligno é maior em pacientes com nível baixo de histiócitos-linfócitos do que em pacientes com níveis altos.

A Figura 3 apresenta uma breve análise de diagnóstico do modelo ajustado, considerando o resíduo componente do desvio. Esse resíduo é sugerido na literatura para análise de diagnóstico em MLG's para dados binários (Paula, 2010; Williams, 1984). Por meio da Figura 3, verifica-se que praticamente todos os valores estão no intervalo $(-2,2)$, e também mostra o bom ajuste do modelo, principalmente da correta distribuição assumida para a variável de interesse. Esta análise gráfica deixa clara a adequação do modelo selecionado, sem indícios de má especificação, corroborando com o resultado do teste gradiente apresentado.

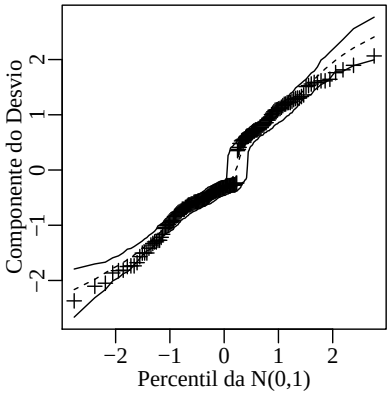


Figura 3: Gráfico de probabilidade normal e envelope simulado dos resíduos componente do desvio.

6 Conclusões

Neste trabalho abordamos o problema da correta especificação da função de ligação em MLG's para dados binários. Propomos um teste RESET baseado na estatística gradiente e comparamos seu desempenho com as estatísticas da razão de verossimilhanças, de Wald e escore. Considerando os resultados numéricos das simulações de Monte Carlo, pode-se observar que, em geral, a estatística G apresenta desempenho equivalente e muitas vezes superior em amostras de tamanho finito. Seu desempenho é verificado em relação a tamanho e poder dos testes, além da distribuição qui-quadrado de referência fornecer uma melhor aproximação para a sua distribuição nula exata.

Indica-se o uso da estatística gradiente para testar a correta especificação da função de ligação em modelos para dados binários, principalmente quando se dispõe de tamanhos amostrais grandes ($n \geq 500$). Além de seu bom desempenho, o cálculo da estatística gradiente é mais simples que a estatística escore, pois não necessita inversão de matrizes e requer menos operações matriciais que a estatística Wald. Como segunda alternativa é indicado o teste de razão de verossimilhanças. A estatística RV também é de fácil utilização, pois é baseada apenas no cálculo das funções de log-verossimilhança maximizadas sob hipóteses nula e alternativa.

Agradecimentos

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

Referências

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–726.

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

Ayalew, L., Yamagishi, H. (2005). The application of gis-based logistic regression for landslide susceptibility mapping in the kakuda-yahiko mountains, central japan. *Geomorphology*, *65*, 15–31.

Bayer, F. M., Cribari-Neto, F. (2013). Bartlett corrections in beta regression models. *Journal of Statistical Planning and Inference*, *143*, 531–547.

Buse, A. (1982). The likelihood ratio, wald and lagrange multiplier tests: An expository note. *The American Statistician*, *Vol. 3*, 153–157.

Chen, H. Z., Randallb, A. (1997). Semi-nonparametric estimation of binary response models with an application to natural resource valuation. *Journal of Econometrics*, *76*, 323–340.

Cordeiro, G. M., Demétrio, C. G. (2007). *Modelos Lineares Generalizados*. Minicurso para o 12º SEAGRO e a 52ª Reunião Anual da RBRAS UFSM, Santa Maria, RS.

Demétrio, C. G. B. (2001). *Modelos Lineares Generalizados em Experimentação Agronômica*. 46ª Reunião Anual da RBRAS e 9º SEAGRO.

Harrell-Jr., F. E., Lee, K. L., Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics*, *15*, 361–387.

Hinkley, D. V. (1985). Transformation diagnostics for linear models. *Biometrika*, *72*, 487–496.

Koenker, R., Yoon, J. (2009). Parametric links for binary choice models: A fisherian-bayesian colloquy. *Journal of Econometrics*, pp. 120–130.

Lemonte, A. J. (2013). On the gradient statistic under model misspecification. *Statistics & Probability Letters*, *83*(1), 390–398.

Lemonte, A. J., Ferrari, S. L. P. (2012). Local power and size properties of the lr, wald, score and gradient tests in dispersion models. *Statistical Methodology*, *9*, 537–554.

McCullagh, P., Nelder, J. (1989). *Generalized linear models*, 2º edn. Chapman and Hall.

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, *135*, 370–384.

Neyman, J., Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *20*, 175–240.

Oliveira, J. S. C. (2013). Detectando má especificação em regressão beta. Dissertação de Mestrado, Universidade Federal de Pernanbuco.

Paula, G. A. (2010). *Modelos de Regressão com Apoio Computacional*. Editora: IME-USP.

Pereira, T. L., Cribari-Neto, F. (2013). Detecting model misspecification in inflated beta regressions. *Communications in Statistics - Simulation and Computation*, *43*, 631–656.

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

Ramalho, E. A., Ramalho, J. J. S. (2012). Alternative versions of the reset test for binary response index models: A comparative study. *Oxford bulletin of economics and statistics*, *74*, 107–130.

Ramsey, J. B. (1969). Tests for specification errorsin classical linear least-squares regression analysis. *Journal of the Royal Statistical Society*, *Vol. 31*, 350–371.

Rao, C. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*(01), 50–57.

Terrell, G. (2002). The gradient statistic. *Computing Science and Statistics*, *34*, 206–215.

Vargas, T. M., Ferrari, S. L. P., Lemonte, A. J. (2013). Gradient statistic: Higher-order asymptotics and Bartlett-type correction. *Electronic Journal of Statistics*, *7*, 43–61.

Vargas, T. M., Ferrari, S. L., Lemonte, A. J. (2014). Improved likelihood inference in generalized linear models. *Computational Statistics and Data Analysis*, *74*, 110–124.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *64*, 426–482.

Wasserman, S., Pattison, P. (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs and*p*. *Psycometrika*, *61*, 401–425.

Williams, D. A. (1984). Residuals in generalized linear models. Em: *International Biometrics Conference*, pp. 59–68.