Comportamento das estatísticas dos testes da Razão de Verssomilhança, Wald e Escore em amostras distintas para a distribuição Beta reparametrizada.

Rosangela Getirana Santana[1], Wesley Oliveira Furriel[2] e André Felipe B. Menezes[3]

[1,2,3] Universidade Estadual de Maringá

RESUMO

Sabendo-se a importância dos indicadores sociais e da saúde para compreensão dos fenômenos que cercam a sociedade contemporânea e a flexibilidade da distribuição de probabilidade Beta para sua modelagem. Este trabalho buscou averiguar as estatísticas dos da Razão de Verossimilhança, Wald e Escore, analisando a probabilidade do *Erro do Tipo I*, com base na condução de um estudo de simulação Monte Carlo, visando avaliá-los em distintos cenários, variando o tamanho das amostras (n) e os valores dos parâmetros μ e ϕ. Para a execução desse procedimento foi adotada uma reparametrização da Beta proposta por Ferrari e Cribari-Neto (2004) presente na análise de regressão. Espera-se obter diferenças entre os testes em relação ao α, quando o tamanho da amostra é pequeno.

Palavras chave: Distribuição Beta, simulação Monte Carlo, tamanho do teste, teste da Razão de Verossimilhança, teste de Wald, teste Escore.

1 Introdução

Nas mais variadas áreas do conhecimento o emprego de indicadores, proporções ou taxas tem sido amplamente utilizado para a sumarização, identificação e hierarquização de distintos fenômenos analisados pelos pesquisadores. Isso ocorre devido ao fato de que, variáveis que assumem um intervalo definido entre um valor mínimo e um máximo, permitem diagnósticos rápidos a cerca de determinado evento, uma

vez que, elas podem, por exemplo, indicar o quão adequado ou inadequado determinado caso está na medida em que ele se aproxima dos valores presentes em umas destas extremidades.

Para melhor compreender essa ideia podemos citar alguns exemplos de variáveis que assumem essas características, como: o IDH (Índice de Desenvolvimento Humano), Cobertura Vacinal, Taxa de Mortalidade, Coeficiente de Gini entre outros.

A família de distribuições Beta é apresentada como uma alternativa na modelagem de variáveis com características descritas acima. Uma variável aleatória X que segue distribuição Beta tem função densidade de probabilidade indexada por dois parâmetros dada por (CASELLA e BERGER, 2002):

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{1}$$

sendo $0 < x < 1$ e α e β ambos parâmetros de forma positivos. A esperança e variância de X são respectivamente dadas por:

$$\mathbb{E}(X) = \frac{\alpha}{(\alpha+\beta)} \text{ e } \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{2}$$

Apesar do intervalo da distribuição ser definido entre 0 e 1, é possível realizar uma generalização dessa característica para um Y que esteja restrito a um intervalo finito (m, n), nos permitindo assim, modelar um conjunto de valores que estejam restritos a um intervalo identificável.

Ferrari e Cribari-Neto(2004) propõe uma reparametrização da distribuição Beta (1) para sua utilização no modelo de regressão. Já que, na análise de regressão, geralmente é mais útil modelar resposta média, então, a fim de obter uma estrutura de regressão para tal, juntamente com um parâmetro de dispersão, é interessante trabalhar com uma parametrização diferente da densidade beta original. O modelo segue as propriedades da Beta, sendo então, adequado para casos em que a variável resposta Y é medida continuamente no intervalo $0 < Y < 1$. Seguindo a nova parametrização, $\mu = \alpha/(\alpha+\beta)$ e $\phi = \alpha+\beta$, ou seja, $\alpha = \mu\,\phi$ e $\beta = (1-\mu)\,\phi$. Isto posto, a média e a variância de Y são dadas pelas seguintes expressões,

$$\mathbb{E}(Y) = \mu \;\text{ e }\; \mathrm{Var}(Y) = \frac{\mathrm{V}(\mu)}{(1+\phi)} \tag{3}$$

em que $\mathrm{V}(\mu) = \mu(1-\mu)$. Tendo em vista essa nova parametrização em termos de μ e ϕ, Y tem-se uma função de densidade Beta dada por:

$$f(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\; 0 < y < 1 \tag{4}$$

com $0 < \mu < 1$ e $\phi > 0$. Assim como na parametrização original é possível obter diferentes formas no comportamento da distribuição de acordo com os valores de seus parâmetros μ e ϕ. Ferrari e Cribari-Neto(2004) apontam que a distribuição pode ser simétrica quando $\mu = 1/2$ e assimétrica quando $\mu \neq 1/2$. Além disso, a dispersão da distribuição, para um μ fixado diminui quando os valores de ϕ aumentam.

Tendo em vista o panorama apresentado o presente trabalho teve como objetivo avaliar os resultados dos testes de hipótese de Razão de Verossimilhança, Wald

e Escore, a partir da condução de um estudo de simulação Monte Carlo, que nos permitiu avaliar os testes, em distintos cenários, nos quais foram variados o tamanho das amostras (n) e os valores dos parâmetros μ e ϕ.

Os três testes citados acima são fundamentados na teoria da verossimilhança, especificamente eles avaliam a função log-verossimilhança em diferentes escalas. A estatística da Razão de Verossimilhança compara a altura das log-verossimilhança dos modelos completo e restrito, ao passo que a estatística de Wald compara a estimativa do parâmetro ($\widehat{\boldsymbol{\theta}}$) com o valor do parâmetro sob a hipótese nula ($\boldsymbol{\theta}_0$). Enquanto que, a estatística Escore verifica a inclinação da log-verossimilhança sob a hipótese nula.

Assim sendo, considere $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ uma amostra aleatória de n observações oriundas de (4), têm-se que o logaritmo natural da função verossimilhança pode ser escrito na forma:

$$\begin{aligned} l(\mu, \phi \mid \boldsymbol{y}) = & n \log\left(\Gamma(\phi)\right) - n \log\left(\Gamma(\mu\phi)\right) - n \log\left(\Gamma((1-\mu)\,\phi)\right) \\ & + (\mu\,\phi - 1) \sum_{i=1}^{n} y_i + [(1-\mu)\phi - 1] \sum_{i=1}^{n} (1 - y_i) \end{aligned} \tag{5}$$

Sabe-se que as estimativas de máxima verossimilhança são os valores de $\boldsymbol{\theta} = (\mu, \phi)$ que maximizam a função log-verossimilhança da amostra (5). Fazendo com que os dados observados sejam tão prováveis quanto possíveis. Desse modo, os valores de $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\phi})$ são as raízes da função escore, dada por:

$$\begin{aligned} \mathcal{U}_{\mu}(\boldsymbol{\theta} \mid \boldsymbol{y}) &= -n\,\phi\,\Psi\left(\mu\,\phi\right) + n\,\phi\,\Psi\left(\phi - \phi\,\mu\right) + \phi \sum_{i=1}^{n} y_i - \phi \sum_{i=1}^{n} \left(1 - y_i\right), \\ \mathcal{U}_{\phi}(\boldsymbol{\theta} \mid \boldsymbol{y}) &= n\,\Psi\left(\phi\right) - n\,\mu\,\Psi\left(\mu\,\phi\right) - n\,\left(1-\mu\right)\,\Psi\left(\phi - \phi\,\mu\right) + \mu \sum_{i=1}^{n} y_i + (1-\mu)\left(n - \sum_{i=1}^{n} y_i\right). \end{aligned}$$

em que Ψ é a função digama. A matriz de informação esperada de $\boldsymbol{\theta}$ através da amostra $\boldsymbol{y}$, é definida como o negativo do valor esperado da segunda derivada da função log-verossimilhança, assim temos:

$$\mathcal{I}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \begin{bmatrix} K_{\mu\,\mu} & K_{\mu\,\phi} \\ K_{\phi\,\mu} & K_{\phi\,\phi} \end{bmatrix} \tag{6}$$

em que:

- $K_{\mu\,\mu} = n\,\phi^2\,\Psi_1\left(\mu\,\phi\right) + n\,\phi^2\,\Psi_1\left(\phi - \phi\,\mu\right)$
- $K_{\mu\,\phi} = K_{\phi\,\mu} = n\,\left[1 - \mu\,\phi\right]\,\Psi_1(\mu\,\phi) - n\,\left[\phi(1-\mu) + 1\right]\,\Psi_1\,\left(\phi - \phi\,\mu\right) + 2n(1-\mu)$
- $K_{\phi\,\phi} = n\,\mu^2\,\Psi_1\left(\mu\,\phi\right) + n\,\left(1-\mu\right)^2\,\Psi_1\left(\phi - \phi\,\mu\right) - n\,\Psi_1\left(\phi\right)$

  sendo Ψ_1 a função trigama.

Para a melhor sistematização e organização da discussão e dos resultados, o trabalho foi dividido em quatro seções. Na seção 2 foram apresentados os testes de hipótese que buscamos avaliar, na seção 3 nos ocupamos em detalhar os cenários sob os quais foi realizado o estudo de simulação e por fim na seção 4, foram apresentados

e discutidos os resultados obtidos. É preciso ressaltar que este estudo tem como intuito iniciar a discussão acerca do uso da distribuição Beta para a modelagem de indicadores sociais e da saúde, atentando-se a sua importância para a compreensão da realidade social e consequentemente para o auxílio das políticas públicas, uma vez que, pretende-se em estudos posteriores realizar uma análise de regressão tendo como variável de interesse o IDH, bem como, indicadores básicos de saúde.

2 Teste de hipóteses

Nesta seção apresentamos a teoria dos testes que foram avaliados no estudo de simulação. Seja $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ uma amostra aleatória de n observações independentes proveniente de (4), suponha o interesse em testar a hipótese nula $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus a alternativa $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, em que $\boldsymbol{\theta} = (\mu, \phi)$.

Através do teste da razão de verossimilhança deve-se avaliar o máximo das funções log-verossimilhanças dos modelos restrito e completo. Sob a hipótese nula a estatística:

$$T_1 = 2\left[l(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{y}) - l(\boldsymbol{\theta}_0 \mid \boldsymbol{y})\right] \tag{7}$$

possui distribuição assintótica qui-quadrado com $p = 2$ graus de liberdade (PAWITAN, 2001).

Por outro lado, o teste de Wald consiste em comparar a estimativa do parâmetro ($\widehat{\boldsymbol{\theta}}$) com o valor do parâmetro sob a hipótese nula ($\boldsymbol{\theta}_0$), por isso requer somente o ajuste do modelo completo. Como temos duas restrições, sob a hipótese nula a estatística do teste tem forma quadrática definida por (MILLAR, 2011):

$$T_2 = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T \mathcal{I}(\widehat{\boldsymbol{\theta}}) \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \tag{8}$$

sendo $\mathcal{I}(\widehat{\boldsymbol{\theta}}) = \mathrm{Var}(\widehat{\boldsymbol{\theta}})$ a matriz de informação esperada localmente nas estimativas de máxima verossimilhança $\widehat{\mu}$ e $\widehat{\phi}$. Desse modo, sob a hipótese nula T_2 tem assintoticamente distribuição qui-quadrado com $p = 2$ graus de liberdade.

Por fim, o teste Escore baseia-se na conjectura da função escore $\mathcal{U}(\boldsymbol{\theta}_0)$ e da matriz de informação esperada $\mathcal{I}(\widehat{\boldsymbol{\theta}}_0)$ sob a hipótese nula. Conforme Millar (2011), a estatística do teste pode ser determinada em termos de forma quadrática por:

$$T_3 = \left(\mathcal{U}(\boldsymbol{\theta}_0)\right)^T \mathcal{I}(\boldsymbol{\theta}_0)^{-1} \mathcal{U}(\boldsymbol{\theta}_0) \tag{9}$$

em que, $\mathcal{U}(\boldsymbol{\theta}_0) = \dfrac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta} \mid \boldsymbol{y})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ e $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$ é a inversa da matriz de informação esperada sob a hipótese nula. A estatística T_3, sob a hipótese nula tem assintoticamente distribuição qui-quadrado com $p = 2$ graus de liberdade. Nota-se que a estatística não requer a maximização da função log-verossimilhança em relação ao modelo completo, exigindo assim menor esforço computacional.

Os autores Mazuchelli e Louzada (2014) ressaltam que estes três testes são assintoticamente equivalentes, todavia podem diferir em amostras pequenas.

3 Estudo de simulação

Diante dos testes descritos na seção anterior, que avaliam as hipóteses por meio de escalas diferentes é de grande interesse verificarmos seu comportamento em distintas situações. Dessa forma, foi conduzido um estudo de simulação Monte Carlo para averiguar o comportamento do tamanho do teste (*Erro do Tipo I*) em diferentes cenários. Assim, foram tomadas amostradas de tamanhos $n = 20, 40, 60, 80$ e 100, $\mu = 0.05, 0.25, 0.5, 0.75$ e 0.95 e $\phi = 5, 15$ e 50. Para cada combinação (n, μ, ϕ) foram geradas $B = 10000$ amostras pseudoaleatórias da nova parametrização da distribuição Beta (4). Em todas as instâncias foram adotados níveis de significância nominais $\alpha = 0,01$ e $\alpha = 0,05$.

Para atingir os objetivos propostos foi implementado no *software* SAS uma *macro* para calcular as probabilidades de rejeitar H_0 quando a mesma é verdadeira (valores-p) dos testes da Seção 2. Sendo que, o teste da Razão de Verossimilhança e Wald foram obtidos via `PROC NLMIXED`, um pacote de procedimentos do SAS que utiliza métodos baseado em verossimilhança para o estudo de modelos não lineares e mistos. Já o teste Escore foi implementado no SAS/IML, pois nenhum procedimento existente nos pacotes do SAS o oferece da forma desejada.

Referências

[1] CASELLA, G.; BERGER, R. L. Statistical inference. Pacific Grove, CA: Duxbury, 2002.

[2] FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. Journal of Applied Statistics, v. 31, n. 7, p. 799-815, 2004.

[3] MAZUCHELI, J.; LOUZADA, F. Discriminação entre as distribuições Odd Weibull e Weibull. Rev. Bras. Biom 32.2 (2014): 226-237.

[4] MILLAR, R. B. Maximum likelihood estimation and inference: with examples in R, SAS and ADMB. Vol. 111. John Wiley & Sons, 2011.

[5] PAWITAN, Y. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, 2001.