REGRESSÃO LOGITO-NORMAL

Brian Alvarez Ribeiro De Melo
Carlos Alberto Ribeiro Diniz

Departamento de Estatística - Universidade Federal de São Carlos

Resumo

Neste trabalho propomos o modelo de regressão logito-normal, uma alternativa ao modelo de regressão beta, com a variável resposta também restrita ao intervalo (0, 1) e seguindo distribuição logito-normal. A média e a variância da distribuição logito-normal não possuem formas fechadas e, por esta razão, tais parâmetros são determinados através de funções dos parâmetros da distribuição Normal μ e σ^2. Estimação por máxima verossimilhança é utilizada para estimar os coeficientes da regressão.

1 Introdução

Modelos de regressão linear são usados para relacionar uma variável a um conjunto de covariáveis. No entanto, estes modelos não são apropriados para situações nas quais a variável resposta está dentro do intervalo unitário, $[0, 1]$. Nestes casos, uma alternativa é transformar a variável resposta para que a mesma assuma valores neste intervalo. Esta alternativa apresenta algumas desvantagens como, por exemplo, o fato de que os parâmetros não são facilmente interpretados. Um modelo de regressão considerado nestes casos é o modelo de regressão beta. O objetivo deste trabalho é apresentar um modelo de regressão alternativo ao modelo de regressão beta, no qual a variável resposta assume valores dentro do intervalo $[0, 1]$, e segue distribuição logito-normal.

Uma caracterização analítica da família de distribuições logito-normal apareceu pela primeira vez no artigo de Johnson (1949). As características dessa família de distribuições ainda são desconhecidas por muitos praticantes da estatística, embora venha sendo usada em modelagem bayesiana hierárquica já há algum tempo.

Enquanto a transformação logito é muito usada para variáveis no intervalo unitário a distribuição logito-normal sempre escapou de exposição direta, o que é de certa forma surpreendente, considerando à difusão atual da análise de regressão logística. Johnson (Johnson, 1949) destaca que enquanto os momentos dessa família de distribuições tem forma muito complicada, o seu cálculo numérico é bem simples.

Uma variável aleatória Y segue distribuição logito normal, dentro do intervalo $[0, 1]$, se a transformação logito, $log[Y/(1-Y)]$, segue distribuição Normal(μ, σ^2). Se $Y \sim LogitoNormal(\mu, \sigma^2)$ então sua função densidade é dada por:

$$f(y) = \frac{1}{y(1-y)\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{log\frac{y}{1-y} - \mu}{\sigma}\right)^2\right\}. \qquad (1)$$

A função de densidade em (1) é simétrica apenas quando $\mu = 0$, e nesse caso $E[Y] = 1/2$. Os valores extremos para a variância de uma distribuição logito-normal, com média M, são 0 e $M(1-M)$, portanto o limite superior da variância coincide com a distribuição Bernoulli.

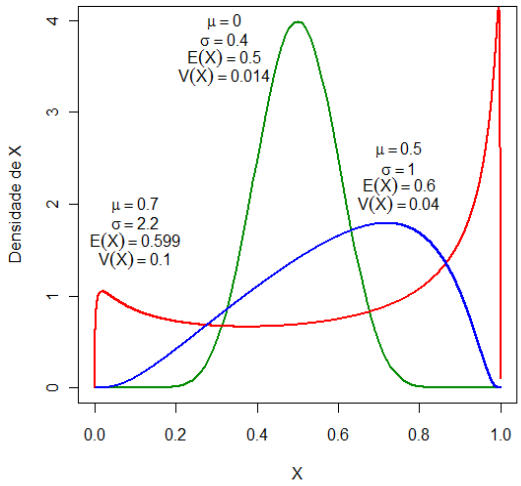

Figura 1: Densidades da logito normal para diferentes valores de μ e σ^2

A distribuição logito-normal é muito versátil, e a sua função de densidade pode adquirir diferentes formas de acordo com os valores dos parâmetros,

μ e σ^2 . Conforme variamos os valores de σ^2, a distribuição logito-normal pode ser unimodal ou bimodal, sendo que nos casos em que σ^2 é grande, a distribuição é bimodal. A média desta distribuição é maior que 0,5 quando μ for maior que 0 e menor que 0,5 sempre que μ for negativo.

Embora não exista forma analítica para os momentos da distribuição logito-normal, sabemos que a esperança e a variância são funções dos parâmetros desta distribuição. Definimos então $E[Y] = H(\frac{\mu}{\sigma}, \sigma)$ e $Var(Y) = V(\frac{\mu}{\sigma}, \sigma)$, onde H e V são funções definidas no intervalo $[0, 1]$.

A Figura 1 apresenta as formas de algumas densidades logito-normal para diferentes valores de μ e $\sigma 2$. Observe que a distribuição que possui o maior valor de σ, e também possui a maior variância, é bimodal, e quando $\mu = 0$ o gráfico é simétrico em torno de 0,5. A Figura 2 mostra a forma da função H, uma aproximação da média da logito-normal,em função de $\frac{\mu}{\sigma}$ e σ.

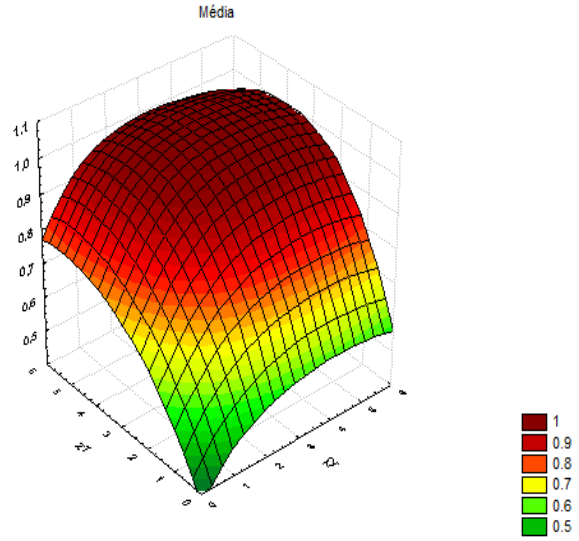

Figura 2: Esperança de Y em função de $\frac{\mu}{\sigma}$ e σ^2

2 Modelo de Regressão

Suponha que $Y_1, Y_2, ..., Y_n$ sejam n variáveis aleatórias independentes, e $Y_i \sim LogitoNormal(\mu_i, \sigma_i^2)$, $i = 1, 2, ..., n$. Então a função de verossimilhança é dada por

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma^2}|y_1, y_2, .., y_n) = \prod_{i=1}^{n} \frac{1}{y_i(1-y_i)\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{1}{2}\left(\frac{log\frac{y_i}{1-y_i} - \mu_i}{\sigma_i} \right)^2 \right\}$$

$$= \frac{(2\pi\sigma_i^2)^{-n/2}}{\prod_{i=1}^{n} y_i(1-y_i)} \exp\left\{ -\frac{1}{2\sigma_i^2}\left[\sum_{i=1}^{n}\left(log\frac{y_i}{1-y_i} \right)^2 - 2\sum_{i=1}^{n}\mu_i log\frac{y_i}{1-y_i} + \sum_{i=1}^{n}\mu_i^2 \right] \right\},$$

com $\boldsymbol{\mu} = (\mu_1, \mu_2, .., \mu_n)'$ e $\boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2, .., \sigma_n^2)'$.

Na construção do modelo de regressão logito-normal os parâmetros μ e σ^2 devem ser escritos como funções da média e da variância da distribuição logito-normal. No entanto, nesta distribuição, os momentos não têm forma fechada, isto é, não existe uma solução analítica para as integrais que representam a esperança e a variância desta distribuição. Por outro lado, a média e a variância da distribuição logito-normal são funções de $\frac{\mu}{\sigma}$ e σ. Para facilitar a notação utilizaremos uma reparametrização na qual consideraremos $\theta_1 = \frac{\mu}{\sigma}$ e $\theta_2 = \sigma$. As funções H e V são determinadas através de procedimentos de simulação, e são dadas por

$$E(Y_i) = H(\frac{\mu_i}{\sigma_i}, \sigma_i) = \gamma_i = (0,20689 + 0,12416\theta_{1i} + 0,0675\theta_{2i})^{\frac{1}{3}}, \qquad (2)$$

$$V(Y_i) = V(\frac{\mu_i}{\sigma_i}, \sigma_i) = \phi = (0,55292 - 0,10333\theta_{1i} + 0,00189\theta_{2i})^4. \qquad (3)$$

Reescrevendo os parâmetros θ_{1i} e θ_{2i} como funções da média γ_i e da variância ϕ, considerando ϕ fixo para todo $i = 1, 2, ..n$ temos

$$\theta_{1i} = -9,4\phi^{\frac{1}{4}} + 0,25\gamma_i^3 + 6,19 \qquad (4)$$

$$\theta_{2i} = 19,23\phi^{\frac{1}{4}} - 15,96\gamma_i^3 - 14 \qquad (5)$$

Substituindo as equações em (4) e (5) na função de verossimilhança encontramos a seguite expressão

$$L(\boldsymbol{\gamma}, \phi) = \prod_{i=1}^{n} \frac{(19,23\phi^{\frac{1}{4}} - 15,96\gamma_i^3 - 14)^{-1}}{y_i(1-y_i)\sqrt{2\pi}} \times$$

$$\times exp\left\{-\frac{1}{2}\left(\frac{log\frac{y_i}{1-y_i}}{19,23\phi^{\frac{1}{4}}-15,96\gamma_i^3-14}-(-9,4\phi^{\frac{1}{4}}+0,25\gamma_i^3+6,19)\right)^2\right\}$$

com $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,...,\gamma_n)$ e ϕ constante.

Assumindo que a média γ_i pode ser escrita como $g(\gamma_i)=\sum_{j=1}^{k}x_{ij}\beta_j$, os parâmetros desconhecidos $\beta_1,\beta_2,...,\beta_k$ são os coeficientes de regressão e g é uma função de ligação monótona e diferenciável.

Considerando g a função de ligação logito, a média γ_i pode ser escrita como

$$\gamma_i=\frac{e^{\sum_{j=1}^{k}x_{ij}}}{1+e^{\sum_{j=1}^{k}x_{ij}}} \tag{6}$$

onde x_{ij} representa o valor da j-ésima covariável para a i-ésima observação.

Os estimadores dos parâmetros de regressão são encontrados maximizando o valor da função de máxima verossimilhança, através de um processo de otimização.

Referências

[1] Ferrari, Silvia L. P.; Cribari-Neto, F.(2003). Beta Regression for modelling rates and proportions.

[2] Lad, F.; Frederic, P.. Two Moments of the Logitnormal Distribution.

[3] Lad, F.; Frederic, P.. A Technical Note on the Logitnormal Distribution.

[4] Johnson, N.L.. Systems of frequency curves generated by methods of translation, Biometrika, 36, pp 149 176, 1949.