XLIII Encontro de Iniciação Científica

Modelo de regressão gama inversa com zeros ajustados

Marcus Vinicius Oliveira Souza [1]
marcusvinicius977.mvo@gmail.com
Manoel Ferreira dos Santos Neto [1]
Rafael Braz de Azevedo Farias[1] Francisco Luan Rodrigues de Sousa[1] Jaiany Nunes Moura[1]
[1] *Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará*

Introdução

A análise de dados com a presença de observações iguais a zero pode representar um desafio em áreas como econometria, ciências biológicas e ciências atuariais, onde variáveis de interesse frequentemente assumem valores nulos em parte significativa das observações. Esse fenômeno é comum, por exemplo, em estudos de frequência de sinistros em seguros de veículos, onde alguns clientes não registram sinistros no período de análise, ou em ecologia animal, onde uma espécie pode não ser observada em certas amostras. Tais dados apresentam uma característica de distribuição que impede o uso direto de modelos de regressão tradicionais, como o modelo de regressão gama, que assume exclusivamente valores positivos na variável dependente.

Na literatura, uma das soluções para lidar com a presença de zeros é o uso de modelos de regressão com zeros ajustados, conhecidos como modelos inflacionados ou aumentados de zeros. Esses modelos tratam separadamente a probabilidade de uma observação ser zero ou positiva, o que permite uma melhor adaptação ao comportamento dos dados.

Recentemente, Vitorino (2024) propôs uma abordagem inovadora ao introduzir a distribuição gama inversa com zeros ajustados. Essa distribuição oferece uma alternativa para modelar dados com zeros, especialmente em situações onde a parte não nula dos dados se ajusta à distribuição gama inversa. No entanto, o trabalho de Vitorino (2024) não considerou uma estrutura de regressão, limitando seu uso a uma análise descritiva dos dados com zeros ajustados.

Neste trabalho, propomos um avanço na metodologia, estendendo a distribuição gama inversa com zeros ajustados para uma estrutura de regressão, de forma a permitir a análise de variáveis dependentes com fatores explicativos associados. O modelo de regressão com zeros ajustados proposto aqui baseia-se na distribuição de Vitorino (2024), mas incorpora uma estrutura de regressão que permitirá avaliar o efeito de covariáveis sobre a variável dependente, oferecendo assim uma ferramenta estatística mais completa e flexível para áreas que trabalham com dados dessa natureza.

Objetivos

Os objetivos deste trabalho são apresentar o modelo de regressão gama inversa com zeros ajustados e realizar um pequeno estudo de simulação para verificar o comportamento dos estimadores de máxima verossimilhança dos coeficientes do modelo.

Distribuição gama inversa com zeros ajustados

A distribuição gama inversa com zeros ajustados (GAIZA) foi proposta por Vitorino (2024). Esta distribuição é a mistura de uma distribuição de Bernoulli e uma distribuição gama inversa. Sua função densidade de probabilidade (FDP) pode ser expressa como:

$$f_Y(y) = \left\{ \frac{(1-p)[\mu(1+\phi)]^{(\phi+2)} \mathrm{e}^{-\frac{\mu(1+\phi)}{y}}}{y^{\phi+3}\Gamma(\phi+2)} \right\}^{1-\rho} \times p^{\rho},$$

em que $\rho = I(y=0), y \geq 0, \phi > 0, \mu > 0$ e $0 < p < 1$.

Modelo de regressão

Definição do modelo

Seja $Y_1, \ldots, Y_n$ uma amostra aleatória em que cada $Y_i \sim \text{GAIZA}(\mu_i, \phi_i, p_i)$ para $i = 1, \ldots, n$. O modelo de regressão GAIZA é definido pelas seguintes relações funcionais:

$$\begin{aligned} g_1(\mu_i) &= \eta_1 = \mathbf{X}_1\beta_1, \\ g_2(\sigma_i) &= \eta_2 = \mathbf{X}_2\beta_2, \\ g_3(p_i) &= \eta_3 = \mathbf{X}_3\beta_3, \end{aligned}$$

em que $g_k(\cdot)$ é uma função de ligação estritamente monótona para $k = 1, 2, 3$, η_k são vetores n-dimensionais, $\mathbf{X}_k$ representam as matrizes de covariáveis de dimensão $n \times J'_k$ e β_k são os vetores de parâmetros de comprimento J'_k.

Estimação dos parâmetros

O logaritmo da função de verossimilhança do modelo de regressão GAIZA é dado por:

$$\begin{aligned} \mathbf{L}(\theta|\mathbf{y}) &= \sum_{y_i \in B_0} \ell(\theta|y_i) + \sum_{y_i \in B_+} \ell(\theta|y_i) \\ &= \sum_{i=1}^{n_0} \log(p_i) + \sum_{i=1}^{n_+} \log(1-p_i) \\ &\quad + (\phi_i + 2) \sum_{i=1}^{n_+} \log(\mu_i(1+\phi_i)) \\ &\quad - (\phi_i + 3) \sum_{i=1}^{n_+} \log(y_i) - \mu(1+\phi_i) \sum_{i=1}^{n_+} \frac{1}{y_i} \\ &\quad - \sum_{i=1}^{n_+} \log(\Gamma(\phi_i + 2)). \end{aligned}$$

em que $B_0 = \{y_i \in \mathbf{y} : y_i = 0\}$, $B_+ = \{y_i \in \mathbf{y} : y_i > 0\}$, $n_0 = n(B_0)$ e $n_+ = n(B_+)$. Para estimar os parâmetros do modelo proposto, colocamos a distribuição GAIZA na estrutura das distribuições mistas do pacote `gamlss` (RIGBY; STASINOPOULOS, 2005). O código está disponível no GitHub e pode ser acessado através do link: https://github.com/statlab-oficial/ZAIGA/blob/main/ZAIGA.R.

Simulação

Neste estudo, realizamos uma simulação de Monte Carlo para avaliar o desempenho dos estimadores dos coeficientes do modelo de regressão GAIZA. Para o ajuste do modelo, utilizou-se a função `gamlss()` do pacote `gamlss` (RIGBY; STASINOPOULOS, 2005). Foram considerados os seguintes cenários para as simulações: tamanhos das amostras $n \in \{50, 100, 200, 400, 600, 800, 1000\}$ e $\beta_{13} \in \{0,2, 0,5, 0,7\}$. Além disso, consideramos 1.000 réplicas de Monte Carlo. As relações funcionais utilizadas foram:

$$\begin{aligned} \log(\mu_i) &= 0,5 + 1,0x_{i2} + 2,5x_{i3}, \\ \log(\sigma_i) &= 1,1 + 2,0z_{i2}, \quad \text{e} \\ p_i &= \beta_{13}, \ (i = 1, 2, \ldots, n), \end{aligned}$$

em que x_{i2}, x_{i3} e z_{i2} foram geradas a partir de distribuições uniformes no intervalo (0,1).

Para cada combinação, geramos dados sintéticos e ajustamos o modelo para estimar os coeficientes. Analisamos o comportamento das estimativas usando métricas como viés relativo (VR) e raiz do erro quadrático médio relativo (REQMR).

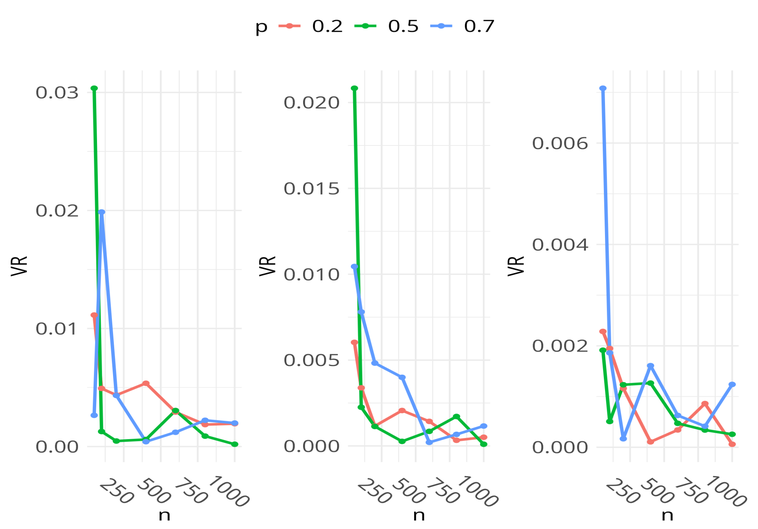

Figure 1: Viés relativo dos estimadores de β_{11} (esquerda), β_{12} (centro) e β_{13} (direita).

Na Figura 1 nota-se que o VR tende a diminuir com o aumento do tamanho da amostra.

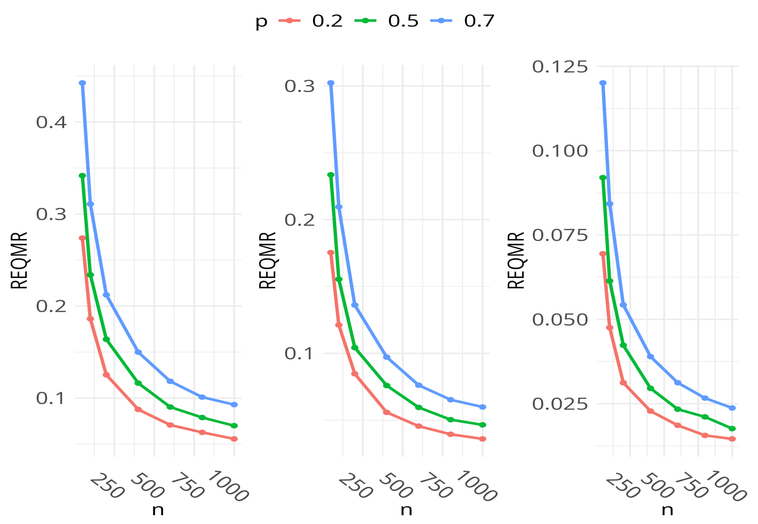

Figure 2: Raiz do erro quadrático médio relativo dos estimadores de β_{11} (esquerda), β_{12} (centro) e β_{13} (direita).

Na Figura 2 é possível notar que os estimadores obtidos na presença de poucos zeros apresentam um desempenho superior em termos de REQMR.

Conclusões

Nota-se, por exemplo, nos resultados das simulações, que o aumento da proporção de zeros reduz a precisão dos estimadores dos coeficientes do modelo.

Além disso, destacamos que um artigo relacionado a este trabalho está em fase de conclusão. A equipe atualmente trabalha na revisão dos resultados e na redação da aplicação do modelo a um conjunto de dados reais.

Referências

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

VITORINO, Rafaella Santos. A distribuição gama inversa com zeros ajustados. 2024. 35 f. Dissertação (Mestrado em Matemática) – Universidade Federal de Campina Grande, Campina Grande, 2024.

Agradecimentos

Este trabalho contou com apoio financeiro parcial da Fundação de Apoio à Pesquisa do Estado da Paraíba - FAPESQ