Modelo de regressão gama inversa com zeros ajustados

Francisco Luan Rodrigues De Sousa[1] **Jaiany Nunes Moura**[1]
Marcus Vinicius Oliveira Souza[1] **Manoel Ferreira dos Santos Neto**[1]

[1] **Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará; luansousa@protonmail.com, jaynunesm@gmail.com, marcusvinicius977.mvo@gmail.com, santosneto@dema.ufc.br**

1. Introdução

A análise de dados com a presença de observações iguais a zero pode representar um desafio em áreas como econometria, ciências biológicas e seguros. Para modelar dados com esta característica, uma parte positiva e outra parte composta por zeros, não se pode utilizar de maneira direta os modelos de regressão tradicionais, como, por exemplo, o modelo de regressão gama. Na literatura, existem os modelos de regressão com zeros ajustados (inflacionados ou aumentados). Recentemente, Vitorino (2024) propôs a distribuição gama inversa com zeros ajustados, porém não trabalhou com estrutura de regressão. Neste trabalho, será proposto um modelo de regressão com zeros ajustados baseado na distribuição proposta por Vitorino (2024).

2. 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.

3. 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:

$$f_Y(y|\mu,\phi,p) = \left\{ \frac{(1-p)[\mu(1+\phi)]^{(\phi+2)} y^{-\phi-3} \exp\left\{-\frac{\mu(1+\phi)}{y}\right\}}{\Gamma(\phi+2)} \right\}^{1-I_{\{0\}}(y)} \times p^{I_{\{0\}}(y)},$$

em que $y \geq 0, \phi > 0, \mu > 0$ e $0 < p < 1$.

4. Modelo de regressão

4.1 Definição 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:

$$g_1(\mu_i) = \sum_{i=1}^{n} \mathbf{x}_i^\top \beta_{1i} = \eta_{\mu,i},$$

$$g_2(\phi_i) = \sum_{i=1}^{n} \mathbf{z}_i^\top \beta_{2i} = \eta_{\phi,i},$$

$$g_3(p_i) = \sum_{i=1}^{n} \mathbf{w}_i^\top \beta_{3i} = \eta_{p,i},$$

em que:

- $\mathbf{x}_i$, $\mathbf{z}_i$ e $\mathbf{w}_i$ são vetores de covariáveis para a i-ésima observação.
- β_{1i}, β_{2i} e β_{3i} são os vetores de coeficientes desconhecidos.

4.2 Estimação dos parâmetros

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

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

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`.

5. Simulações

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 $n \in \{50,65,80,95,110,\ldots,185,200\}$ e $p \in \{0,20,0,50,0,70\}$. Além disso, consideramos 5.000 réplicas de Monte Carlo. As relações funcionais utilizadas foram: $\log(\mu_i) = 0,5 + 1,0x_{i2} + 2,5x_{i3}$, $\log(\sigma_i) = 1,1 + 2,0z_{i2}$, e $\text{logit}(p_i) = \beta_{31}$, $(i = 1,2,\ldots,n)$, 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).

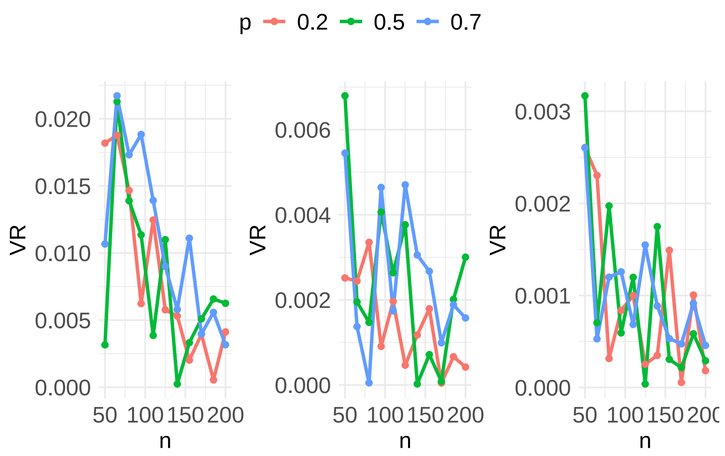

Figura 1: *Viés relativo dos estimadores de β_{11} (esquerda), β_{12} (centro) e β_{13} (direita).*

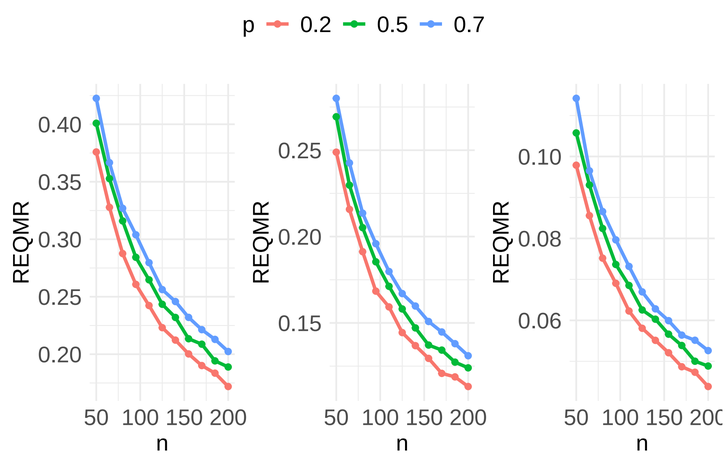

Figura 2: *Raiz do erro quadrático médio relativo dos estimadores de β_{11} (esquerda), β_{12} (centro) e β_{13} (direita).*

6. 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.

7. Referências bibliográficas

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.

8. Agradecimentos

Este trabalho contou com o financiamento da Fundação de Apoio à Pesquisa do Estado da Paraíba - FAPESQ.