trab

Giovanni Samartini, Luís Porto, Vinicius Bronzatti, Vinicius Oliveira

18/07/2020

Introdução

A regressão Binomial Negativa Inflado em Zeros (ZINB) é utilzada para dados de contagem que apresentam superdispersão e excesso de zeros, de maneira que a distribuição dos dados é de forma combinada entre uma distribuição de Binomial Negativa e uma função Logito.

Objetivo

O objetivo do trabalho é aplicar o Modelo de Regressão Binomial Negativa Inflado em Zeros em um banco de dados e comparar sua eficácia com o Modelo de Regressão Binomial Negativa comum.

Metodologia

Suponha que para cada observação existem dois casos possíveis. Suponha que se o caso 1 ocorre, a contagem é 0 e se o caso 2 ocorre, as contagens (incluindo os 0) são geradas de acordo com o modelo de binomial negativo. Se o caso 1 ocorre com probabilidade π e o caso 2 ocorre com probabilidade $1-\pi$, a distribuição de probabilidades de uma variável aleatória ZINB, dada por y_i, pode ser escrita da forma:

$$P(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)(g(y_i = 0), & \text{se j=0} \\ (1-\pi_i)g(y_i), & \text{se j>0} \end{cases}$$

Onde π_i é a função de ligação logística definida abaixo e $g(y_i)$ é a distribuição da binomial negativa definida por:

$$g(y_i) = P(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

A componente da binomial negativa pode incluir um tempo de exposição t e um conjunto de k variáveis regressoras (os x's). A expressão em relação a essas quantidades se dá por:

$$\mu_i = exp(ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki})$$

Na maioria dos casos, $x_1 \equiv 1$, então β_1 é chamado de intercepto. Os coeficientes de regressão $\beta_1, \beta_2, ..., \beta_k$ são parâmetros desconhecidos e estimados pelo conjunto de dados.

A função de ligação logística π_i é dada por:

$$\pi_i = \frac{\lambda_i}{1+\lambda_i}$$

Onde

$$\lambda_i = exp(ln(ti) + \gamma_1 z_{1i} + \gamma_2 z_{2i} + ... + \gamma_m z_{mi}$$

O componente logístico inclui um tempo de exposição t e um conjunto de m variáveis regressoras (os z's). Note que os z's e os x's podem ou não ter termos em comum.

Aplicação

Ajustando um modelo para modelar a contagem de peixes pegos utilizando as variáveis *child* e *camper* para modelar a contagem e a variável *persons* na parte logito, temos:

Na Tabela 1 temos o sumário das variáveis da Regressão Binomial Negativa, enquanto na Tabela 3, temos o sumário das variáveis do Modelo de Inflação, que inclui os coeficientes da Logito para predizer excesso de zeros.

Tabela 1: Estimativas, Erros Padrão, Valores Z e p-valores dos parâmetros da Regressão Binomial Negativa.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.3710	0.2561	5.3533	0.0000
child	-1.5153	0.1956	-7.7470	0.0000
camper1	0.8791	0.2693	3.2645	0.0011
Log(theta)	-0.9854	0.1760	-5.6002	0.0000

Tabela 2: Estimativas, Erros Padrão, Valores Z e p-valores dos parâmetros do Modelo de Inflação.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6031	0.8365	1.9164	0.0553
persons	-1.6666	0.6793	-2.4534	0.0142

Temos que todos os preditores em ambas porções do modelo são estatisticamente significantes, sugerindo um bom ajuste aos dados. Para confirmar isso, vamos comparar esse modelo com o modelo padrão de Binomial Negativa, que resultou no sumário mostrado na Tabela 3.

Tabela 3: Estimativas, Erros Padrão, Valores Z e p-valores dos parâmetros do Modelo Padrão de Binomial Negativa.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.0727	0.2425	4.4239	0.0000
child	-1.3753	0.1958	-7.0251	0.0000
camper1	0.9094	0.2836	3.2060	0.0013

Nesse caso, também temos que todos os preditores são estatisticamente significantes. Para determinar qual o melhor modelo dentre os dois, será aplicado o teste de Vuong, para verificar qual modelo se aproxima mais dos dados originais.

No teste de Vuong, obtemos a estatística Z = 1.701 e um p-valor = 0.044. Com isso, rejeitamos a hipótese nula de que os modelos são indiferenciáveis e concluimos que o modelo inflacionado em zeros significativamente melhor que o modelo padrão.

Bibliografia

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Zero-Inflated_Negative_Binomial_Regression.pdf