Modelos de Regressão de Poisson Inflacionados em Zero

ME714 | ANÁLISE DE DADOS DISCRETOS
Profa. Dra. Hildete Prisco Pinheiro

Caroline da Silva Mangile 195539
Gabriela Inocente Yogi 141812
Rodrigo Resende Soares Rocha 186819
Wesley R. da Silva Satelis 188650

2 de julho de 2021

1 Introdução

ainda tá bem ruim

Neste trabalho são expostos conceitos teóricos a respeito dos modelos de Poisson Inflacionados em Zero (ZIP) e uma aplicação utilizando um conjunto de dados real. O conjunto é formado por viagens de acampamento feitas por 250 grupos de pessoas à um parque nos Estados Unidos.

Foram feitas análises descritivas, diagnósticos de modelo, interpretações a respeito do problema e predição. Todo o trabalho foi conduzido com o uso da linguagem e ambiente de computação estatística R (R Core Team 2021).

2 Métodos

2.1 Distribuição de Poisson

2.2 Teste de superdispersão

O teste de superdispersão proposto por Cameron and Trivedi (1990) é baseado em uma regressão linear sem o intercepto com as hipóteses

$$H_0 : Var\left(y_i\right) = \lambda_i \quad vs \quad H_1 : Var\left(y_i\right) = \lambda_i + \alpha g\left(\lambda_i\right)$$

em que α é um parâmetro desconhecido e $g(.)$ é uma função definida, comumente $g(\lambda_i) = \lambda_i^2$ ou $g(\lambda_i) = \lambda_i$. Este teste é condusido estimando-se o modelo de Poisson, construindo $\widehat{\lambda_i} = \exp\left(\mathrm{x}_i'\widehat{\beta}\right)$ e ajustando um modelo por mínimos quadrados ordinários sem o intercepto

$$\frac{(y_i - \widehat{\lambda_i})^2 - y_i}{\widehat{\lambda_i}} = \alpha\frac{g(\widehat{\lambda_i})}{\widehat{\lambda_i}} + e_i$$

em que e_i é o erro. A significância do coeficiente α implica na existencia de superdispersão nos dados.

2.3 Modelos de Poisson Inflacionados em Zero (ZIP)

Uma propriedade importante da distribuição de Poisson é que a média e variância são iguais, $Var(y_i|x_i) = E(y_i|x_i) = \lambda_i$, esta propriedade é referida como equidispersão. Na prática a suposição de equidispersão é violada quando a variância das contagens observadas é maior que a média por conta de heterogeneidade não observada ou quando a frequência de zeros é maior que o número de zeros esperado em uma distribuição de Poisson.

Assumindo que a variável resposta tem distribuição de Poissson e que o logaritmo de seu valor esperado pode ser modelado por uma combinação linear de parâmentros desconhecidos. Seja y_i, $i = 1, ..., n$ a variável resposta de um modelo de regressão, assumimos que y_i tem distribuição de Poisson com média λ_i, definida em função das covariáveis x_i. Assim, um modelo de regressão de Poisson é dado por

$$P(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

em que a esperança condicional é dado por $\lambda_i = E(y_i|x_i) = \exp(x'\beta)$. O vetor $\mathrm{x}_i' = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ contem as covariáveis e $\beta' = (\beta_1, ..., \beta_p)$ é o vetor de parâmetros do modelo de regressão.

A regressão de Poisson é inadequada quando temos excesso de zeros na amostra porque viola a suposição de equidispersão. Lambert (1992) introduziu o modelo de Poisson Inflacionado em Zero (ZIP) como uma alternativa na modelagem de dados deste tipo. Ele assume que as respostas provêm de dois processos. Um processo modela inflações em zero, incluindo uma proporção $1-p$ de zeros extras e uma $p\exp(\lambda_i)$ de zeros da distribuição de Poisson; e o segundo modela as contagens diferentes de zero usando um modelos de Poisson trucado em zero. Assim, o modelo ZIP é dado por

$$\mathrm{P}\left(\mathrm{Y_i} = \mathrm{y_i} \mid \mathrm{x_i}, \mathrm{z_i}\right) = \begin{cases} \theta_{\mathrm{i}}\left(z_{\mathrm{i}}\right) + \left(1 - \theta_{\mathrm{i}}\left(z_{\mathrm{i}}\right)\right)\mathrm{Pois}\left(\lambda_{\mathrm{i}}; 0 \mid \mathrm{x_i}\right) & \text{if} \quad \mathrm{y_i} = 0 \\ \left(1 - \theta_{\mathrm{i}}\left(\mathrm{z_i}\right)\right)\mathrm{Pois}\left(\lambda_{\mathrm{i}}; \mathrm{y_i} \mid \mathrm{x_i}\right) & \text{if} \quad \mathrm{y_i} > 0 \end{cases}$$

em que z_i é um vetor de covariáveis definindo a probabilidade θ_i. A média e variância do modelo são $(1-\theta_i)\lambda_i$ e $(1-\theta_i)(\lambda_i + \theta_i\lambda_i^2)$, respectivamente.

2.4 Teste de Voung

3 Aplicação

Os dados são provenientes de 250 acampamentos familiares em um parque nos Estados Unidos. Cada grupo foi questionado sobre o número de peixes capturados, quantas pessoas e quantas crianças o grupo tinha, e se eles foram acompanhados por um guia.

- **LIVE_BAIT:** Variável binária. Indica se foram usadas iscas vivas ou não;
- **CAMPER:** Variável binária. Indica se o o grupo foi acompanhado por um guia ou não;
- **PERSONS:** Variável numérica. Número de pessoas no grupo;
- **CHILDREN:** Variável numérica. Número de crianças no grupo;
- **FISH_COUNT:** Variável numérica. Número de peixes pegos pelo grupo;

Pelo histograma da Figura 1, nota-se um número elevado de zeros, o que é um indicativo de que um modelo ZIP provavelmente se ajuste melhor aos dados.

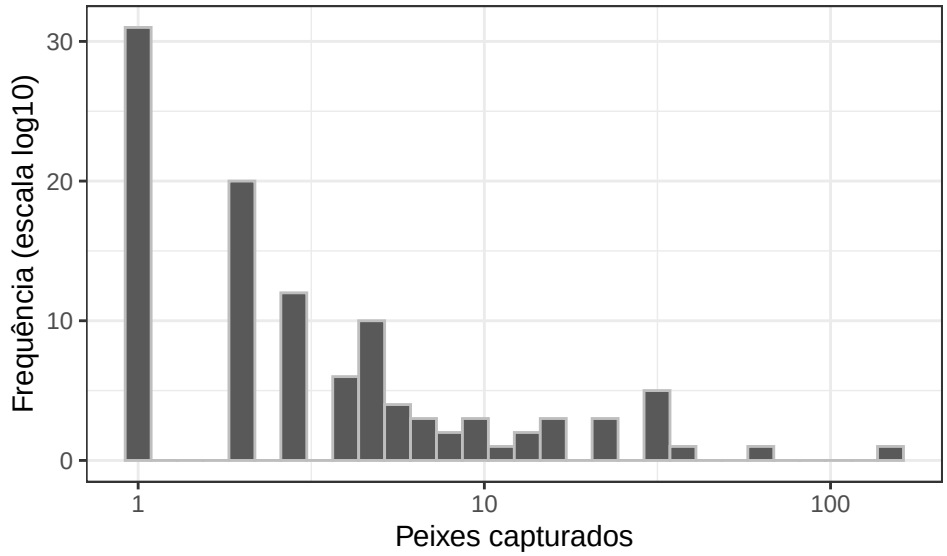

Figure 1: Histograma de frequências do número de peixes capturados (variável resposta).

Utilizando o teste de superdispersão proposto por Cameron and Trivedi (1990), chegamos à conclusão de que existe superdispersão nos dados ($z \approx 1,8573,\ \alpha \approx 10,6243,\ p-valor \approx 0,03163$). Como a suposição de equidispersão não é satisfeita, não podemos ajustar um modelo de regressão de Poisson padrão. Pelos resultado do teste de Voung e da Figura 1, temos indícios que a superdispersão é causada pelo excesso de zeros na variável resposta (`FISH_COUNT`).

Ajustando um modelo ZIP com as variáveis explicativas `CHILDREN` e `CAMPER` no modelo de contagem e `PERSONS` no modelo de zeros, chegamos aos resultados apresentados na Tabela 1.

Table 1: Nome da tabela.

Estimate	Std.Error	Z.Value	P.Valor	Parametro	Porcento2.5	Porcento97.5
1,5979	0,0855	18,6804	<0.001	Intercepto	1,4302	1,7655
-1,0428	0,1000	-10,4296	<0.001	CHILDREN	-1,2388	-0,8469
0,8340	0,0936	8,9079	<0.001	CAMPER (1)	0,6505	1,0175
1,2974	0,3739	3,4705	5e-04	Intercepto	0,5647	2,0302
-0,5643	0,1630	-3,4630	5e-04	PEARSONS	-0,8838	-0,2449

Podemos ainda nos perguntar se o modelo ajustado realmente é superior à regressão padrão de Poisson. Comparando os dois ajustes com o teste de Voung chegamos a um $p-valor \approx 0,00018$, evidenciando que o modelo ZIP é superior ao modelo de Poisson padrão.

Com o teste qui-quadrado baseado no log das verossimilhanças **definido em...** podemos testar a hipótese que o modelo ZIP ajustado é superior ao modelo somente com o intercepto. Com um p-valor $< 0,0001$, temos evidências para afirmar que o modelo ajustado é superior ao modelo somente com intercepto.

incident risk ratio (IRR) for the Poisson model and odds ratio (OR) for the logistic (zero inflation) model.

4 Conclusões

Referências

Cameron, A.Colin, and Pravin K. Trivedi. 1990. "Regression-Based Tests for Overdispersion in the Poisson Model." *Journal of Econometrics* 46 (3): 347–64. https://doi.org/https://doi.org/10.1016/0304-4076(90)90014-K.

Lambert, Diane. 1992. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics* 34 (1): 1–14.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.