Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1806-2563
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v35i2.15071

A study of the score test in discrimination poisson and zero-inflated poisson models

Vanessa Siqueira Peres da Silva[1], Marcelo Ângelo Cirillo[2*] and Juliana Garcia Cespedes[3]

*[1]Departamento de Ciências Exatas, Universidade Federal de Lavras, Campus Universitário, Lavras, Minas Gerais, Brazil. [2]Departamento de Ciências Exatas, Universidade Federal de Lavras, Cx. Postal 3037, 37200-000, Lavras, Minas Gerais, Brazil. [3]Departamento de Ciência e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, São Paulo, Brazil. *Author for correspondence. E-mail: macufla@dex.ufla.br*

ABSTRACT. In many experimental situations the sample may present excess zero observations and generally are used probabilistic models for zero inflated to represent them. However no one knows precisely the amount of zero observations that these models support. Depending on the sample size and null observations number the Poisson model can be used. Based on this question, the objective of this paper is to evaluate the properties of Type I error and power of the score test (proposed by Van Den Broek (1995) to discriminate the Poisson and Zero-inflated Poisson models) and ascertain the most appropriate model to represent a sample with excess zeros without compromising the statistical inference. Through Monte Carlo simulation we concluded that when considering a sample of size at least n = 40 with 30% of the null observations, the score test had a high discriminatory power between the ZIP and Poisson model indicating that in fact is relevant the use of the ZIP model.

Keywords: zero-inflated Poisson, score test, Monte Carlo simulation.

Um estudo do teste de escore na discriminação de modelos poisson e poisson inflacionados de zeros

RESUMO. Em inúmeras situações experimentais a amostra pode apresentar excesso de observações iguais a zero e geralmente utilizam-se modelos probabilísticos inflacionados de zero para representá-las. Contudo não se sabe com precisão a quantidade de observações nulas que esses modelos suportam. Dependendo do tamanho amostral e do número de observações nulas o modelo de Poisson pode ser utilizado. Tendo por base essa questão, o objetivo desse trabalho é avaliar as propriedades do erro tipo I e poder do teste de escore (proposto por Van Den Broek (1995) para discriminar os modelos Poisson e Poisson Inflacionado de zero) e verificar qual é o modelo mais adequado para representar uma amostra com excesso de zeros sem comprometer a inferência estatística. Por meio de simulação Monte Carlo concluiu-se que ao considerar uma amostra de tamanho no mínimo n = 40 contendo 30% de observações nulas o teste de escore apresentou um alto poder discriminatório entre o modelo Poisson e ZIP indicando que de fato é pertinente o uso do modelo ZIP.

Palavras-chave: Poisson inflacionada de zeros, teste de escore, simulação Monte Carlo.

Introduction

Before defining which probabilistic model is used in data modeling, the researcher may be faced with samples that have a significant amount of zero observations. In this situation, the usual method of parameters estimation, the maximum likelihood method, should be modified so that estimates can contemplate the effect of the presence of these observations.

The excess of zero observations is also a root causes of outliers, Copas (1988) warned that, even using the appropriate model, a number of observations can be detected as outlier because the value of the π ratio is close to 0 or 1. The author also points out that the maximum likelihood estimation is sensitive to the presence of outliers and suggests a correction in the estimates because different models used for response have varied sensitivity in detecting outliers. To work around this problem, zero-inflated models can be used in the analysis of these data (CZADO et al., 2007; FAMOYE; SINGH, 2006; LAMBERT, 1992; MIN; CZADO, 2010; MULLAHAY, 1997; SLYMEN et al., 2006; SILVA; CIRILLO, 2010).

Some questions arise upon the foregoing: If we know a certain amount of zeros observations in the sample, are the usual maximum likelihood estimates reliable? Referring this issue to count data: can be trusted to use the Poisson model since it is known the number of zero observations that this model

supports? Note that, if we assume the usual maximum likelihood estimates, the expected variability in the model may be greater than the sample variance, thus resulting in the over dispersion effect.

So, given this problem, to know the effect of zero observations in the parameter estimates is essential for the researcher to secure it in statistical inference to be performed.

The test proposed by Van Den Broek (1995) considers H_0: $p = 0$ versus H_1: $p \neq 0$, where p is the proportion of zeros contained in the population. The rejection of the null hypothesis of this test leads us to interpret the model to be used should be the zero inflated Poisson (ZIP) model. Evidently, the effect of zero observations in the ZIP model is built to estimate the average rate, as well as the expected variance in the model.

In this context, other inferential procedures have been proposed. As an example we can mention the likelihood ratio test (EL-SHAARAWI, 1985) and confidence intervals for p (XIE et al., 2001). Importantly, these tests are asymptotic, so when you have small samples the reliability of statistical results can be inaccurate (DENG; PAUL, 2005; GUPTA, et al., 2004; JANSAKUL; HINDE, 2002). On this regard, we emphasize the motivation of this article in order to study the control of type I error and power of the test, particularly the score test proposed by Van Den Broek (1995) in zero inflated samples. For this purpose a Monte Carlo simulation study was carried out considering different configurations for sample sizes and parameter values described in the next section.

Material and methods

Let Y be a random variable with probability function given by (JOHNSON et al., 1992):

$$f(Y,p,\theta)=\begin{cases} p+(1-p)\exp(-\theta), & \text{se } y=0 \\ (1-p)\exp(-\theta)\dfrac{\theta^{y}}{y!}, & \text{se } y>0 \end{cases}, \quad (1)$$

where:

p refers to the proportion of zeros;

θ is the average rate of the Poisson distribution;

$f(Y,p,\theta)$ refers to zero-inflated Poisson distribution.

On the null hypothesis $p = 0$, the function (1) is reduced to the Poisson distribution.

A random sample of size n $\{y_1, y_2,...,y_n\}$ has been generated from the model (1) and the number of observations equal to zero was defined by n_0. For the remaining sample values $y = 1, 2, ...$ we used the notations n_1 to number of observations equal to 1, n_2 for the number of observations equal to 2 and so on, so that $n = n_0 + \sum_{y=1}^{\infty} n_y$. Thus, the likelihood function is defined in Equation (2):

$$\begin{aligned} L(p,\theta\,|\,y_i) &= \prod_{i=1}^{n}\left[p+(1-p)\exp(-\theta)I_{\{y_i=0\}}+(1-p)\exp(-\theta)\frac{\theta^{y_i}}{y_i!}I_{\{y_i>0\}}\right] \\ &= [p+(1-p)\exp(-\theta)]^{n_0}\left[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}\right]^{n-n_0} \end{aligned} \quad (2)$$

The first order partial derivatives were calculated using the log-likelihood function, denoted by $l(p,\theta\,|\,y_i)$, and are described by expressions (3) and (4):

$$\frac{dl(p,\theta\,|\,y_i)}{dp}=\frac{n_0(1-\exp(-\theta))}{p+(1-p)\exp(-\theta)}-\frac{(n-n_0)}{(1-p)} \quad (3)$$

$$\frac{dl(p,\theta\,|\,y_i)}{d\theta}=\frac{n_0(1-p)\exp(-\theta)}{p+(1-p)\exp(-\theta)}+(n-n_0)\frac{(y_i-\theta)}{\theta} \quad (4)$$

The second order partial derivatives were obtained using the Newton-Raphson method to obtain the maximum likelihood estimation, following the iterative process given by Equation (5):

$$\tilde{\beta}^{(i+1)}=\tilde{\beta}^{(i)}-H_{\tilde{\beta}}^{-1}\nabla_l, \quad (5)$$

where $H_{\tilde{\beta}}^{-1}$ and ∇_l are the Hessian matrix and the gradient vector, respectively defined by (6) and (7):

$$H_{\tilde{\beta}}^{-1}=\begin{bmatrix} \dfrac{dl^2(p,\theta\,|\,y_i)}{d^2\theta} & \dfrac{dl^2(p,\theta\,|\,y_i)}{d\theta dp} \\ \dfrac{dl^2(p,\theta\,|\,y_i)}{dpd\theta} & \dfrac{dl^2(p,\theta\,|\,y_i)}{d^2p} \end{bmatrix} \quad (6)$$

$$\nabla_l=\left[\frac{dl(p,\theta\,|\,y_i)}{dp},\frac{dl(p,\theta\,|\,y_i)}{d\theta}\right] \quad (7)$$

Obtained the maximum likelihood estimates of the ZIP model, we proceeded with the evaluation of the score test, whose statistic defined under H_0: $p = 0$ is given by Equation (8) that asymptotically converges to the chi-square distribution with one degree of freedom.

$$S=\frac{(n_0-np_0)^2}{np_0(1-p_0)-n\bar{y}p_0^2} \quad (8)$$

where n is the sample size, the mean of observations is represented by $\bar{y}$ and n_0 is the number of zero observations contained in the sample. Considering the maximum likelihood estimate of the Poisson parameter obtained on the null hypothesis, the proportion of null values is specified by $p_0 = \exp(-\hat{\theta})$. Thus, the rejection of H_0 implies that the ZIP model should be used to adjust the data distribution.

By the above, using the Monte Carlo simulation, it was carried out an assessment of the properties of the control of type I error and power of the test score so that the likelihood of Type I error was determined by the proportion of times where the statistic S (8) was rejected at nominal level of significance set at 5% when the samples were simulated on H_0. Similarly the power was computed, however, considering the simulated samples of H_1. Thus, the parametric values used in 10,000 Monte Carlo iterations were specified as the following description: (a) Poisson distribution average rate *(θ)* fixed at 5 and 10; (b) Different parametric values set by p = 1, 3, 5, 10 and 30%; (c) Different sample sizes *(n)* set at n = 25, 30, 40, 50, 100, 150, 200, 500, 1000, 5000 and 10000.

Finally, to obtain the results we implemented a program in the software R Development Core Team (2011), using the package ZIGP. Thus, the effective control of type I error provided by the test, considering a sample size that expresses high value of power, which is allowed to recommend the required sample size for the score test can be used in the discrimination of Poisson or ZIP models.

Results and discussion

Occurrence rate of type I error of score test.

The results discussion for the type I error control was made in a comparative way, considering different values of the average rate *(θ)*. Thus, as described in Table 1, we were observed that the increase of the Poisson average rate from the $\theta = 5$ to $\theta = 10$ led to different results regarding the occurrence rate of Type I error, and, in the initial $(\theta = 5)$ the probability of type I error was close to the nominal level of significance from samples $n \geq 40$. For $\theta = 10$ the results were conservative for almost all sample sizes.

This fact is somewhat consistent with the construction of the score test proposed by Van Den Broek (1995), as besides the test is asymptotic, the distribution of score statistic is approximately distributed by chi-square, and therefore, necessarily it requires assumption of normality. In the case of the Poisson model $(H_0: p = 0)$, a known result is that this assumption is verified for high average rates.

Table 1. Type I error rates for the scores test as a function of sample size *(n)*, average rate Poisson *(θ)*, nominal level of significance $\alpha = 5\%$ and $N = 10{,}000$ simulations.

N	*θ = 5*	*θ = 10*
25	0.0890	0.0011
30	0.0750	0.0015
40	0.0480	0.0016
50	0.0470	0.0023
100	0.0508	0.0044
150	0.0460	0.0068
200	0.0475	0.0094
500	0.0499	0.0280
1,000	0.0553	0.0250
5,000	0.0578	0.0215
10,000	0.0572	0.0570

In accordance with the results related to control Type I error (Table 1) proceeded with the discussion of results for the test power, as described in Section 3.2.

Score test power

Keeping the same parametric specifications regarding the Poisson average rate of $(\theta = 5$ *and* $\theta = 10)$ and different sample sizes, the score test power was evaluated under the alternative hypothesis considering the proportions of zero values hypothesized by H_1 *at* $p = 1, 3, 5, 10$ *and* 30%.

The values obtained for the study of the test power are described in Table 2. However, it should be emphasized that according to the results on the probability of occurrence of Type I error control have been consistent with the nominal level of significance set at 0.05 in both parametric values assumed by θ, the discussion of results stopped for samples of size n ≥ 40.

The power values obtained for $\theta = 10$ were lower than the initial situation $(\theta = 5)$. However, it was clear the relationship between sample size *(n)* with the proportion of null values *(p)* in the approximation of results from power, considering both parametric values of θ assumed. Thus, it was found that, as there is increase in sample size simultaneously with the reduction of p values of power tend to be similar.

Due to this trend, in practical terms, it is appropriate to recommend a sample size to be used by the researcher to provide reliability in the application of the score test, given the knowledge that the average rate of Poisson approaches between the parametric values of $\theta = 5$ and $\theta = 10$ units. Accordingly, the results are described in Table 3.

Table 2. Power of the test score (%) for various values of θ and nominal level of significance *(α = 5%)* depending on the percentage of zeros *(p)* and sample size *(n)*.

(n, p)	$\theta = 5$	$\theta = 10$
(40; 0.01)	33.8	18.4
(40; 0.03)	69.8	47.6
(40; 0.05)	87.1	69.9
(40; 0.10)	98.7	94.1
(40; 0.30)	100.0	100.0
(50; 0.01)	40.2	20.7
(50; 0.03)	78.4	53.7
(50; 0.05)	92.6	76.8
(50; 0.10)	99.4	97.3
(50; 0.30)	100.0	100.0
(100; 0.01)	63.2	27.9
(100; 0.03)	95.2	73.9
(100; 0.05)	99.5	93.3
(100; 0.10)	100.0	99.9
(100; 0.30)	100.0	100.0
(150; 0.01)	77.5	37.1
(150; 0.03)	98.9	86.2
(150; 0.05)	99.9	98.1
(150; 0.10)	100.0	99.9
(150; 0.30)	100.0	100.0
(200; 0.01)	87.1	40.9
(200; 0.03)	99.7	92.5
(200; 0.05)	100.0	99.5
(200; 0.10)	100.0	100.0
(200; 0.30)	100.0	100.0
(500; 0.01)	99.5	67.8
(500; 0.03)	100.0	99.5
(500; 0.05)	100.0	100.0
(500; 0.10)	100.0	100.0
(500; 0.30)	100.0	100.0
(1,000; 0.01)	100.0	89.2
(1,000; 0.03)	100.0	100.0
(1,000; 0.05)	100.0	100.0
(1,000; 0.10)	100.0	100.0
(1,000; 0.30)	100.0	100.0
(5,000; 0.01)	100.0	100.0
(5,000; 0.03)	100.0	100.0
(5,000; 0.05)	100.0	100.0
(5,000; 0.10)	100.0	100.0
(5,000; 0.30)	100.0	100.0
(10,000; 0.01)	100.0	100.0
(10,000; 0.03)	100.0	100.0
(10,000; 0.05)	100.0	100.0
(10,000; 0.10)	100.0	100.0
(10,000; 0.30)	100.0	100.0

Table 3. Smaller sample size on the proportion of zeros that can provide values similar to the average $\theta = 5$ and $\theta = 10$, for score test used to verify the adequacy of the Poisson or ZIP models in a sample with excess of zeros, considering different values of *p*.

p	Sample size (*n*)
30%	n ≥ 40
10%	n ≥ 100
5%	n ≥ 150
3%	n ≥ 200
1%	n ≥ 500

In synthesis the interpretation of results is given in the specification of the suitable sample size to verify the adequacy of the Poisson and ZIP models through score test proposed by Van den Broek. Thus, considering a sample with a ratio of 0.3 for null values, so that the test triggers a high value of power in the indication of use of the Poisson or ZIP models it is recommended $n \geq 40$. If this proportion is 10%, the minimum sample size should be $n \geq 100$. For a proportion of 5% the minimum sample size should be at least $n \geq 150$. Similarly the other recommendations are given considering different sample sizes evaluated.

Conclusion

Through the evaluated scenarios it was found that as there is increase in sample size simultaneously with the reduction of *p,* values of power tend to be similar.

For high zeros proportion assumed *(p = 30%)* there is statistical evidence that a given sample of $n = 40$ the score test proposed by Van den Broek has a high discriminatory power between the Poisson and ZIP models.

References

COPAS, J. B. Binary regression models for contaminated data. **Journal of the Royal Statistical Society: Series B, Methodological**, v. 50, n. 2, p. 225-265, 1988.

CZADO, C.; ERHARDT, V.; MIN. A.; WAGNER, S. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. **Statistical Modelling**, v. 7, n. 2, p. 125-153, 2007.

DENG, D.; PAUL, S. R. Score tests for zero-inflation and over-dispersion in generalized linear models. **Statistica Sinica**, v. 15, n. 2, p. 257-276, 2005.

EL-SHAARAWI, A. H. Some goodness-of-fit methods for the Poisson plus added zeros distribution. **Applied and Environmental Microbiology**, v. 49, n. 5, p. 1304-1306, 1985.

FAMOYE, F.; SINGH, K. Zero-inflated generalized Poisson regression model with an application to domestic violence data. **Journal of Data Science**, v. 4, n. 1, p. 117-130, 2006.

GUPTA, P. L.; GUPTA, R. C.; TRIPATHI, R. C. Score test for zero inflated generalized Poisson regression model. **Communications in Statistics: Theory and Methods**, v. 33, n. 1, p. 47-64, 2004.

JANSAKUL, N.; HINDE, J. P. Score tests for zero-inflated Poisson models. **Computational Statistics Data Analysis**, v. 40, n. 1, p. 75-96, 2002.

JOHNSON, N.; KOTZ, S.; KEMP, A. W. **Univariate discrete distributions**. 2nd ed. New York: J. Wiley, 1992.

LAMBERT, D. Zero-inflated Poisson regression with application to defects in manufacturing. **Technometrics**, v. 34, n. 1, p. 1-14, 1992.

MIN, A.; CZADO, C. Testing for zero-modification in count regression models. **Statistica Sinica**, v. 20, n. 1, p. 323-341, 2010.

MULLAHAY, J. Heterogeneity, excess zeros, and the structure of count data models. **Journal of Applied Econometrics**, v. 12, n. 3, p. 337-350, 1997.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2011.

SILVA, A. M.; CIRILLO, M. A. Estudo por simulação monte carlo de um estimador robust utilizado na inferência de um modelo binomial contaminado. **Acta Scientiarum. Technology**, v. 32, n. 3, p. 303-307, 2010.

SLYMEN, D. J.; AY ALA, G. A.; ARREDONDO, E. M.; ELDER, J. P. A demonstration of modeling count data with an application to physical activity. **Epidemiologic Perspectives and Innovations**, v. 3, n. 3, p. 1-9, 2006.

VAN DEN BROEK, J. A score test for zero inflation in a Poisson-distribution. **Biometrics**, v. 51, n. 2, p. 738-743, 1995.

XIE, M.; HE, B.; GOH, T. N. Zero-inflated Poisson model in statistical process control. **Computing and Statistical Data Analysis**, v. 38, n. 2, p. 191-201, 2001.

Received on October 19, 2011.
Accepted on April 11, 2012.