Statistical Papers
https://doi.org/10.1007/s00362-023-01477-2

REGULAR ARTICLE

Bartlett corrections for zero-adjusted generalized linear models

Tiago M. Magalhães[1] · **Gustavo H. A. Pereira**[2] · **Denise A. Botter**[3] · **Mônica C. Sandoval**[3]

Received: 15 August 2022 / Revised: 3 April 2023

Abstract

Zero-adjusted generalized linear models (ZAGLMs) are used in many areas to fit variables that are discrete at zero and continuous on the positive real numbers. As in other classes of regression models, hypothesis testing inference in the class of ZAGLMs is usually performed using the likelihood ratio statistic. However, the LR test is substantially size distorted when the sample size is small. In this work, we derive an analytical Bartlett correction of the LR statistic. We also consider two different adjustments for the LR statistic based on bootstrap. Monte Carlo simulation studies show that the improved LR tests have null rejection rates close to the nominal levels in small sample sizes and similar power. An application illustrates the usefulness of the improved statistics.

Keywords Bartlett corrections · Chi-squared distribution · Maximum likelihood estimates

Mathematics Subject Classification 62Fxx · 62F05

1 Introduction

Zero-adjusted regression models (ZAR models) are often used to fit variables that are discrete at zero and continuous at some interval of the positive real numbers. They are used in many areas such as insurance (Bortoluzzo et al. 2011), botany (Thomson et al. 2018), credit risk (Tong et al. 2016), microbiology (Rocha et al. 2017), biodiversity

✉ Tiago M. Magalhães
tiago.magalhaes@ufjf.br

1 Federal University of Juiz de Fora, Juiz de Fora, Minas Gerais, Brazil

2 Federal University of São Carlos, São Carlos, São Paulo, Brazil

3 University of São Paulo, São Paulo, São Paulo, Brazil

Published online: 19 September 2023

(Rubec et al. 2016) and meteorology (Zamani and Bazrafshan 2020). ZAR models are also known as zero-augmented regression models (Nogarotto et al. 2020) or zero-inflated regression models. The last expression is used especially when the continuous component of the response variable is beta distributed (Ospina and Ferrari 2012). However, zero-inflated regression models usually refer to models in which the response variable is discrete with more zeros than expected by a known probability distribution (Lambert 1992). Recent works related to ZAR models include Tomazella et al. (2019); Calsavara et al. (2019); Hashimoto et al. (2019); Pereira et al. (2020); Michaelis et al. (2020); Ye et al. (2021) and Silva et al. (2021).

Zero-adjusted generalized linear models (ZAGLMs) are a subclass of ZAR models, in which the continuous component of the regression model is a generalized linear model (Dunn and Smyth 2018). The main components of the class of ZAGLMs are the zero-adjusted gamma regression models (ZAGA regression models, Tong et al. 2013) and the zero-adjusted inverse Gaussian regression models (ZAIG regression models, Heller et al. 2006).

Hypothesis testing inference in the class of ZAGLMs is usually performed using the likelihood ratio (LR) statistic, especially when the null hypothesis of interest involves more than one parameter. Under the null hypothesis, the LR statistic has an asymptotic chi-squared distribution (Sen et al. 2010). However, in many regression models, the chi-squared distribution is not a good approximation of the null distribution of the LR statistic when the sample size is small (Melo et al. 2009; Pereira and Cribari-Neto 2014). As a consequence, in these cases, the test based on the LR statistic is often size distorted.

An alternative to improve the chi-squared approximation to the exact null distribution of the LR statistic is to use the Bartlett correction (Bartlett 1937; Lawley 1956). The Bartlett correction is usually effective in bringing the true sizes of the test of the model closer to the nominal levels (Botter and Cordeiro 1997). In practical situations, type I errors should be nearer to the fixed nominal value (usually 1%, 5% or 10%) than the original statistic. Many authors have presented Bartlett correction factors for specific regression models. Cordeiro (1983) derived the Bartlett correction factor for generalized linear models and Botter and Cordeiro (1997) extended it to double generalized linear models. This correction was derived for mixed linear models and for beta regression by Melo et al. (2009) and Bayer and Cribari-Neto (2013), respectively. Moulton et al. (1993) and Das et al. (2018) showed that the Bartlett correction also improves the LR statistic in logistic regression. Recent works related to this topic include Loose et al. (2018); Araújo et al. (2020); Magalhães and Gallardo (2020); Rauber et al. (2020); Guedes et al. (2020, 2021) and Melo et al. (2022).

Few works have proposed small-sample adjustments to the LR statistic in ZAR models. Pereira and Cribari-Neto (2014) derived a correction for the LR statistic known as Skovgaard's adjustment (Skovgaard 2001) for zero-adjusted beta regressions. For the same model, Loose et al. (2017) proposed a Bartlett correction based on bootstrap (Efron 1979), instead of the traditional analytical correction. To the best of our knowledge, no study has proposed small-sample adjustments to the LR statistic in ZAGLMs. Moreover, the previous works involving ZAR models have not studied the behavior of the adjusted LR statistic when the null hypothesis of interest involves parameters of more than a submodel of the ZAR model. In practice, it is often desirable to test whether the distribution of the response variable is related to a given covariate. In these cases, the null hypothesis has parameters of all ZAGLM submodels.

The chief goal of our paper is to improve the LR statistic in the class of ZAGLMs. Two approaches are used. First, we derive an analytical Bartlett correction of the LR statistic. In addition, we propose using two different adjustments for the LR statistic based on bootstrap (Cordeiro and Cribari-Neto 2014). The performance in small and medium-sized samples of the adjusted statistics is compared with the usual LR statistics through extensive Monte Carlo simulation studies.

The remainder of the paper is organized as follows. Section 2 defines the ZAGLMs and presents some of their inferential aspects. The adjusted LR statistic is derived in Sect. 3 and the bootstrap corrections are also described in that section. In the following section, Monte Carlos simulation studies are performed to compare the finite sample behavior of different LR statistics. Section 5 presents an application to real data. Concluding remarks are provided in Sect. 6.

2 Model

Suppose that the univariate random variable $Y \in \{0\} \cup (0, \infty)$, has a density with the following

$$g(y; \pi, \theta, \phi) = \begin{cases} 0 & \text{if } y < 0, \\ \pi & \text{if } y = 0, \\ (1-\pi) f(y; \theta, \phi) & \text{if } y > 0, \end{cases} \tag{1}$$

where $\pi = \mathbb{P}(Y = 0)$ and $f(y; \theta, \phi)$ is a probability density function (PDF) of a positive continuous random variable. The expression (1) can be written as:

$$g(y; \pi, \theta, \phi) = \left\{ \pi^{\mathbb{I}_{\{0\}}^{(y)}} (1-\pi)^{1-\mathbb{I}_{\{0\}}^{(y)}} \right\} \left\{ f(y; \theta, \phi)^{1-\mathbb{I}_{\{0\}}^{(y)}} \right\}, \tag{2}$$

with

$$\mathbb{I}_{\{0\}}^{(y)} = \begin{cases} 1 \text{ if } y = 0, \\ 0 \text{ otherwise.} \end{cases}$$

In our work, we define

$$f(y, \theta, \phi) = \exp\{\phi[y\theta - b(\theta) + c(y)] + d_1(y) + d_2(\phi)\}, \tag{3}$$

where $b(\cdot)$, $c(\cdot)$, $d_1(\cdot)$ and $d_2(\cdot)$ are known functions, i.e., $f(\cdot, \theta, \phi)$ is the PDF of a member of the exponential family (EF) with parameters θ and ϕ, the canonical and the precision parameters respectively (the inverse, ϕ^{-1}, is the dispersion parameter). If $Z \sim \text{EF}(\theta, \phi)$, then $\mathbb{E}(Z) = db(\theta)/d\theta = \mu$ and $\text{Var}(Z) = \phi^{-1} d^2 b(\theta)/d\theta^2 = \phi^{-1} V(\mu)$, where $V = V(\mu)$ is the variance function. Note that the function of the random variable $\mathbb{I}_{\{0\}}^{(y)}$, in (2), can be seen as a probability function of a Bernoulli distribution with probability π, also a member of the EF. Therefore, $\mathbb{E}(Y) = (1-\pi)\mu$ and $\mathrm{Var}(Y) = (1-\pi)[\pi\mu^2 + \phi^{-1}V(\mu)]$. Table 1 presents the quantities presented in (3) for the distributions in the EF used this paper.

We consider that (1) has three systematic components, which are parameterized as $\mu = \mu(\boldsymbol{\beta})$, $\phi = \phi(\boldsymbol{\delta})$ and $\pi = \pi(\boldsymbol{\gamma})$. The systematic components are:

$$\eta_1 = h_1(\mu) = \boldsymbol{x}^\top\boldsymbol{\beta},\ \eta_2 = h_2(\phi) = \boldsymbol{t}^\top\boldsymbol{\delta},\ \eta_3 = h_3(\pi) = \boldsymbol{s}^\top\boldsymbol{\gamma}, \tag{4}$$

where $h_1(\cdot)$ to $h_3(\cdot)$ are the link functions and are known one-to-one continuously four-times differentiable functions, η_1 to η_3 are the linear predictors, $\boldsymbol{x} = (x_1, \ldots x_{p_\mu})^\top$, $\boldsymbol{t} = (t_1, \ldots t_{p_\phi})^\top$, $\boldsymbol{s} = (s_1, \ldots s_{p_\pi})^\top$ are specified regressor vectors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_1})^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{p_2})^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_3})^\top$ are sets of unknown parameters to be estimated. We also assume that $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are functionally independent from each other.

Consider $Y_1, \ldots, Y_n$ independent random variables from (1) and the parameter vector $\boldsymbol{\lambda} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top)^\top$. The logarithm of the likelihood (log-likelihood) function based on a sample of n independent observations is given by

$$l(\boldsymbol{\lambda}) = l_1(\boldsymbol{\beta}, \boldsymbol{\delta}) + l_2(\boldsymbol{\gamma}), \tag{5}$$

where

$$l_1(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{\ell: y_\ell \in (0,\infty)} \left\{\phi_\ell \left[y_\ell\theta_\ell - b\left(\theta_\ell\right) + c\left(y_\ell\right)\right] + d_1\left(y_\ell\right) + d_2\left(\phi_\ell\right)\right\}$$

and

$$l_2(\boldsymbol{\gamma}) = \sum_{\ell=1}^{n} \left\{\mathbb{I}_{\{0\}}^{(y_\ell)} \log\left(\frac{\pi_\ell}{1-\pi_\ell}\right) + \log(1-\pi_\ell)\right\}.$$

The function (5) is assumed to be regular with respect to all $\boldsymbol{\lambda}$ derivatives. The score vector, obtained by differentiation of the log-likelihood function $l(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$, can be written as $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\lambda}) = \left(\boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\lambda})^\top, \boldsymbol{U}_{\boldsymbol{\delta}}(\boldsymbol{\lambda})^\top, \boldsymbol{U}_{\boldsymbol{\gamma}}(\boldsymbol{\lambda})^\top\right)^\top$, with

$$\boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \boldsymbol{X}^\top\boldsymbol{\Phi}\boldsymbol{W}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{I}_{\boldsymbol{y}}(\boldsymbol{y}-\boldsymbol{\mu}),$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ is a specified $n \times p_1$ matrix of full rank $p_1 < n$, $\boldsymbol{\Phi} = \mathrm{diag}\{\phi_1, \ldots, \phi_n\}$, $\boldsymbol{W} = \mathrm{diag}\{w_1, \ldots, w_n\}$, $w_\ell = (d\mu_\ell/d\eta_{1\ell})^2 V_\ell^{-1}$, $\boldsymbol{V} = \mathrm{diag}\{V_1, \ldots, V_n\}$, $\boldsymbol{I}_{\boldsymbol{y}} = \mathrm{diag}\left\{1-\mathbb{I}_{\{0\}}^{(y_1)}, \ldots, 1-\mathbb{I}_{\{0\}}^{(y_n)}\right\}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$,

$$\boldsymbol{U}_{\boldsymbol{\delta}}(\boldsymbol{\lambda}) = \boldsymbol{T}^\top\boldsymbol{\Phi}_1\boldsymbol{I}_{\boldsymbol{y}}\boldsymbol{v},$$

Table 1 Important positive distributions belonging to the exponential family

Distribution	θ	$b(\theta)$	ϕ	$V(\mu)$	$c(y)$	$d_1(y)$	$d_2(\phi)$
Bernoulli	$\log(\mu/1-\mu)$	$\log(1+e^{\mu})$	1	$\mu(1-\mu)$	0	0	0
Gamma	$-1/\mu$	$-\log\theta$	$1/CV^2$	μ^2	$\log y$	$-\log y$	$\phi\log\phi-\log\Gamma(\phi)$
Inverse Gaussian	$-1/2\mu^2$	$-\sqrt{-2\theta}$	ϕ	μ^3	$-1/2y$	$-(1/2)\log 2\pi y^3$	$(1/2)\log\phi$

where $\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)^\top$ is a specified $n \times p_2$ matrix of full rank $p_2 < n$, $\boldsymbol{\Phi}_1 = \mathrm{diag}\{\phi_{11}, \ldots, \phi_{1n}\}$, $\phi_{1\ell} = d\phi_\ell/d\eta_{2\ell}$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^\top$, $\nu_\ell = y_\ell\theta_\ell - b(\theta_\ell) + c(y_\ell) + dd_2(\phi_\ell)/d\phi_\ell$,

$$\boldsymbol{U}_{\boldsymbol{\gamma}}(\boldsymbol{\lambda}) = \boldsymbol{S}^\top \boldsymbol{W}_\pi^{1/2} \boldsymbol{V}_\pi^{-1/2} (\boldsymbol{y}^I - \boldsymbol{\pi}),$$

where $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)^\top$ is a specified $n \times p_3$ matrix of full rank $p_3 < n$, $\boldsymbol{W}_\pi = \mathrm{diag}\{w_{\pi 1}, \ldots, w_{\pi n}\}$, $w_{\pi\ell} = (d\pi_\ell/d\eta_{3\ell})^2 V_{\pi\ell}^{-1}$, $\boldsymbol{V}_\pi = \mathrm{diag}\{V_{\pi 1}, \ldots, V_{\pi n}\}$, $V_{\pi\ell} = \pi_\ell(1 - \pi_\ell)$, $\boldsymbol{y}^I = (y_1^I, \ldots, y_n^I)^\top$, $y_\ell^I = \mathbb{I}_{\{0\}}^{(y_\ell)}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$. The partition $\boldsymbol{\lambda} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top, \boldsymbol{\gamma}^\top)^\top$ induces a corresponding partitioned Fisher information matrix for these parameters. This matrix is block-diagonal given by:

$$\boldsymbol{K}_{\boldsymbol{\lambda},\boldsymbol{\lambda}} = \begin{pmatrix} \boldsymbol{X}^\top \boldsymbol{\Delta} \boldsymbol{W} \boldsymbol{\Phi} \boldsymbol{X} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{T}^\top \boldsymbol{\Delta} \boldsymbol{D}_2 \boldsymbol{\Phi}_1^2 \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{S}^\top \boldsymbol{W}_\pi \boldsymbol{S} \end{pmatrix}, \qquad (6)$$

where $\boldsymbol{\Delta} = \mathrm{diag}\{1-\pi_1, \ldots, 1-\pi_n\}$, $\boldsymbol{D}_2 = \mathrm{diag}\{d_{21}, \ldots, d_{2n}\}$, $d_{2\ell} = d^2 d_2(\phi_\ell)/d\phi_\ell^2$, $\boldsymbol{\Phi}_1^2 = \mathrm{diag}\{\phi_{11}^2, \ldots, \phi_{1n}^2\}$, $\phi_{1\ell}^2 = (d\phi_\ell/d\eta_{2\ell})^2$. Thus, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$ are globally orthogonal (Cox and Reid 1987) and their maximum likelihood estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ are asymptotically independent. The former property is necessary to simplify the calculations of the Bartlett corrections, whereas the latter is desirable in the context of inference. The Fisher scoring method can be used to compute $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ by iteratively solving the following equations:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{(m+1)} \\ \hat{\boldsymbol{\delta}}^{(m+1)} \\ \hat{\boldsymbol{\gamma}}^{(m+1)} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(m)} \\ \hat{\boldsymbol{\delta}}^{(m)} \\ \hat{\boldsymbol{\gamma}}^{(m)} \end{pmatrix} + \left[\boldsymbol{K}_{\boldsymbol{\lambda},\boldsymbol{\lambda}}^{-1}\right]^{(m)} \times \boldsymbol{U}^{(m)}.$$

In many problems, the restrictions under a test involve a subset of the $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ parameters. We partition the parameters as $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top\right)^\top$, $\boldsymbol{\delta} = \left(\boldsymbol{\delta}_1^\top, \boldsymbol{\delta}_2^\top\right)^\top$ and $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top\right)^\top$ where $\boldsymbol{\beta}_1 = \left(\beta_1, \ldots, \beta_{q_1}\right)^\top$, $\boldsymbol{\beta}_2 = \left(\beta_{q_1+1}, \ldots, \beta_{p_1}\right)^\top$, $\boldsymbol{\delta}_1 = \left(\delta_1, \ldots, \delta_{q_2}\right)^\top$, $\boldsymbol{\delta}_2 = \left(\delta_{q_2+1}, \ldots, \delta_{p_2}\right)^\top$, $\boldsymbol{\gamma}_1 = \left(\gamma_1, \ldots, \gamma_{q_3}\right)^\top$ and $\boldsymbol{\gamma}_2 = \left(\gamma_{q_3+1}, \ldots, \gamma_{p_3}\right)^\top$. The partitions of $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ induce the corresponding partitions $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, $\boldsymbol{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2)$, $\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2)$, $\boldsymbol{U} = \left(\boldsymbol{U}_{\boldsymbol{\beta}_1}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^\top, \boldsymbol{U}_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^\top, \boldsymbol{U}_{\boldsymbol{\delta}_1}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)^\top, \boldsymbol{U}_{\boldsymbol{\delta}_2}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)^\top, \boldsymbol{U}_{\boldsymbol{\gamma}_1}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^\top, \boldsymbol{U}_{\boldsymbol{\gamma}_2}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^\top\right)^\top$ and

$$\boldsymbol{K}_{\boldsymbol{\beta},\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{K}_{\boldsymbol{\beta}_{11}} & \boldsymbol{K}_{\boldsymbol{\beta}_{12}} \\ \boldsymbol{K}_{\boldsymbol{\beta}_{21}} & \boldsymbol{K}_{\boldsymbol{\beta}_{22}} \end{pmatrix}, \ \boldsymbol{K}_{\boldsymbol{\delta},\boldsymbol{\delta}} = \begin{pmatrix} \boldsymbol{K}_{\boldsymbol{\delta}_{11}} & \boldsymbol{K}_{\boldsymbol{\delta}_{12}} \\ \boldsymbol{K}_{\boldsymbol{\delta}_{21}} & \boldsymbol{K}_{\boldsymbol{\delta}_{22}} \end{pmatrix}, \ \boldsymbol{K}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} = \begin{pmatrix} \boldsymbol{K}_{\boldsymbol{\gamma}_{11}} & \boldsymbol{K}_{\boldsymbol{\gamma}_{12}} \\ \boldsymbol{K}_{\boldsymbol{\gamma}_{21}} & \boldsymbol{K}_{\boldsymbol{\gamma}_{22}} \end{pmatrix},$$

where $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are known matrices of full rank and dimensions $n \times q_1$, $n \times (p_1 - q_1)$, $n \times q_2$, $n \times (p_2 - q_2)$, $n \times q_3$, $n \times (p_3 - q_3)$, respectively, and $\boldsymbol{K}_{\boldsymbol{\beta}_{11}} = \boldsymbol{X}_1^\top \boldsymbol{\Delta} \boldsymbol{W} \boldsymbol{\Phi} \boldsymbol{X}_1$, $\boldsymbol{K}_{\boldsymbol{\beta}_{12}} = \boldsymbol{K}_{\boldsymbol{\beta}_{21}}^\top = \boldsymbol{X}_1^\top \boldsymbol{\Delta} \boldsymbol{W} \boldsymbol{\Phi} \boldsymbol{X}_2$, $\boldsymbol{K}_{\boldsymbol{\beta}_{22}} = \boldsymbol{X}_2^\top \boldsymbol{\Delta} \boldsymbol{W} \boldsymbol{\Phi} \boldsymbol{X}_2$, $\boldsymbol{K}_{\boldsymbol{\delta}_{11}} =$

$-\boldsymbol{T}_1^\top \boldsymbol{\Delta} \boldsymbol{D}_2 \boldsymbol{\Phi}_1^2 \boldsymbol{T}_1$, $\boldsymbol{K}_{\delta_{12}} = \boldsymbol{K}_{\delta_{21}}^\top = -\boldsymbol{T}_1^\top \boldsymbol{\Delta} \boldsymbol{D}_2 \boldsymbol{\Phi}_1^2 \boldsymbol{T}_2$, $\boldsymbol{K}_{\delta_{22}} = -\boldsymbol{T}_2^\top \boldsymbol{\Delta} \boldsymbol{D}_2 \boldsymbol{\Phi}_1^2 \boldsymbol{T}_2$, $\boldsymbol{K}_{\gamma_{11}} = \boldsymbol{S}_1^\top \boldsymbol{W}_\pi \boldsymbol{S}_1$, $\boldsymbol{K}_{\gamma_{12}} = \boldsymbol{K}_{\gamma_{21}}^\top = \boldsymbol{S}_1^\top \boldsymbol{W}_\pi \boldsymbol{S}_2$ and $\boldsymbol{K}_{\gamma_{22}} = \boldsymbol{S}_2^\top \boldsymbol{W}_\pi \boldsymbol{S}_2$.

We are interested in testing

$$\begin{cases} \mathcal{H} : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^{(0)}, \boldsymbol{\delta}_1 = \boldsymbol{\delta}_1^{(0)}, \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_1^{(0)} \\ \mathcal{A} : \text{violation of at least one equality} \end{cases} \tag{7}$$

where $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\delta}_1^{(0)}$ and $\boldsymbol{\gamma}_1^{(0)}$ are specified vectors of dimensions q_1, q_2 and q_3, respectively. We assume that $0 \leq q_1 \leq p_1$, $0 \leq q_2 \leq p_2$ and $0 \leq q_3 \leq p_3$, but the trivial case $q_1 = q_2 = q_3 = 0$ is excluded because there are no parameters left under the null hypothesis. Let $\hat{\boldsymbol{\lambda}} = \left(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\delta}}^\top, \hat{\boldsymbol{\gamma}}^\top\right)^\top$ be the unrestricted maximum likelihood estimates of $\boldsymbol{\beta}, \boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ and $\tilde{\boldsymbol{\lambda}} = \left(\boldsymbol{\beta}_1^{(0)\top}, \tilde{\boldsymbol{\beta}}_2^\top, \boldsymbol{\delta}_1^{(0)\top}, \tilde{\boldsymbol{\delta}}_2^\top, \boldsymbol{\gamma}_1^{(0)\top}, \tilde{\boldsymbol{\gamma}}_2^\top\right)^\top$ be their restricted maximum likelihood estimates under $\mathcal{H}$. The likelihood ratio statistic for testing $\mathcal{H}$ is

$$\mathrm{LR} = 2\left[l\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\delta}}_2, \hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2\right) - l\left(\boldsymbol{\beta}_1^{(0)}, \tilde{\boldsymbol{\beta}}_2, \boldsymbol{\delta}_1^{(0)}, \tilde{\boldsymbol{\delta}}_2, \boldsymbol{\gamma}_1^{(0)}, \tilde{\boldsymbol{\gamma}}_2\right)\right],$$

which is, under $\mathcal{H}$ and some regularity conditions, asymptotically distributed as $\chi^2_{q_1+q_2+q_3}$ with approximation error of order n^{-1}.

3 Corrected likelihood ratio tests

It is known that under the null hypothesis and general conditions of regularity, the likelihood ratio statistic, LR, has an asymptotic χ^2_q distribution, where q is the number of restrictions imposed on the parameters by the null hypothesis. Lawley (1956) improved the LR statistic by defining a statistic, say $\mathrm{LR_c}$, such that $\mathbb{P}_{\mathcal{H}}(\mathrm{LR_c} \leq w) = \mathbb{P}\left(\chi^2_q \leq w\right) + \mathcal{O}(n^{-2})$, while $\mathbb{P}_{\mathcal{H}}(\mathrm{LR} \leq w) = \mathbb{P}\left(\chi^2_q \leq w\right) + \mathcal{O}(n^{-1})$, where $\mathbb{P}_{\mathcal{H}}(\cdot)$ is the cumulative distribution function under the null hypothesis. The improved statistic is given by $\mathrm{LR_c} = \tilde{c}^{-1}\mathrm{LR}$, where $\tilde{c} = q^{-1}\tilde{\mathbb{E}}(\mathrm{LR})$ is a consistent estimate of $c = q^{-1}\mathbb{E}(\mathrm{LR})$ and $\mathbb{E}(\mathrm{LR})$ is the expectation of the likelihood ratio statistic, evaluated under the null hypothesis up to order n^{-1}. The factor c is known as the Bartlett correction factor.

For the model presented in Sect. 2, define $q = q_1 + q_2 + q_3$ and $p = p_1 + p_2 + p_3$. The expected likelihood ratio statistic to order $\mathcal{O}(n^{-1})$ for the test of hypotheses in (7) is $\mathbb{E}(\mathrm{LR}) = q + \varepsilon_p - \varepsilon_{p-q}$, where

$$\begin{aligned} \varepsilon_p = {} & \varepsilon_{\beta_{p_1}} + \varepsilon_{\delta_{p_2}} + \varepsilon_{\gamma_{p_3}} \\ & + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Delta}\boldsymbol{\Phi}_2 \boldsymbol{W} \boldsymbol{Z}_{\beta d} \boldsymbol{Z}_{\delta d} - \boldsymbol{\Delta}\boldsymbol{\Phi}_1 \boldsymbol{W}(\boldsymbol{Z}_\beta^{(2)} \odot \boldsymbol{Z}_\delta)\boldsymbol{W}\boldsymbol{\Phi}_1\boldsymbol{\Delta}\right\} \\ & + \frac{1}{4}\mathbf{1}^\top \boldsymbol{\Delta}\boldsymbol{\Phi}_1 \boldsymbol{W} \boldsymbol{Z}_{\beta d} \boldsymbol{Z}_\delta \left[\boldsymbol{Z}_{\beta d}\boldsymbol{W} + 2\boldsymbol{Z}_{\delta d}\left(\boldsymbol{\Phi}_1^2 \boldsymbol{D}_3 + \boldsymbol{\Phi}_2 \boldsymbol{D}_2\right)\right]\boldsymbol{\Phi}_1\boldsymbol{\Delta}\mathbf{1}, \end{aligned} \tag{8}$$

with $\mathbf{1} = (1, \ldots, 1)_{n\times 1}^{\top}$, $\mathbf{\Phi}_2 = \text{diag}\{\phi_{21}, \ldots, \phi_{2n}\}$, $\phi_{2\ell} = d^2\phi_\ell/d\eta_{2\ell}^2$, $\mathbf{Z}_\beta = \mathbf{X}\mathbf{K}^{\boldsymbol{\beta},\boldsymbol{\beta}}\mathbf{X}^\top$, $\mathbf{Z}_{\beta d} = \text{diag}\{z_{\beta,11}, \ldots, z_{\beta,nn}\}$, $\mathbf{Z}_\delta = \mathbf{T}\mathbf{K}^{\boldsymbol{\delta},\boldsymbol{\delta}}\mathbf{T}^\top$, $\mathbf{Z}_{\delta d} = \text{diag}\{z_{\delta,11}, \ldots, z_{\delta,nn}\}$, $\mathbf{Z}_\gamma = \mathbf{S}\mathbf{K}^{\boldsymbol{\gamma},\boldsymbol{\gamma}}\mathbf{S}^\top$, $\mathbf{Z}_{\gamma d} = \text{diag}\{z_{\gamma,11}, \ldots, z_{\gamma,nn}\}$, $\mathbf{D}_3 = \text{diag}\{d_{31}, \ldots, d_{3n}\}$, $d_{3\ell} = d^3 d_2(\phi_\ell)/d\phi_\ell^3$, $\mathbf{Z}_\beta^{(2)} = \mathbf{Z}_\beta \odot \mathbf{Z}_\beta$, $\mathbf{Z}_\beta^{(3)} = \mathbf{Z}_\beta^{(2)} \odot \mathbf{Z}_\beta$, $\mathbf{Z}_\delta^{(2)} = \mathbf{Z}_\delta \odot \mathbf{Z}_\delta$, $\mathbf{Z}_\delta^{(3)} = \mathbf{Z}_\delta^{(2)} \odot \mathbf{Z}_\delta$, $\mathbf{Z}_\gamma^{(2)} = \mathbf{Z}_\gamma \odot \mathbf{Z}_\gamma$, $\mathbf{Z}_\gamma^{(3)} = \mathbf{Z}_\gamma^{(2)} \odot \mathbf{Z}_\gamma$ and $\odot$ represents a direct product of matrices (Hadamard product). For sake of brevity, $\varepsilon_{\beta_{p_1}}, \varepsilon_{\delta_{p_2}}, \varepsilon_{\gamma_{p_3}}$ are given in the Appendix. The term ε_p is of order $\mathcal{O}(n^{-1})$ evaluated at the true parameter point. For sake of brevity, we will not discuss particular cases derived from (8). However, they can be obtained similarly as Botter and Cordeiro (1997).

The Bartlett-corrected LR test statistic for testing (7) is

$$\text{LR}_\text{c} = \frac{\text{LR}}{(1+\zeta)},$$

where $\zeta = \left(\varepsilon_p - \varepsilon_{p-q}\right)/q$, ε_{p-q} can be determined from (8) with $\mathbf{X}_2$, $\mathbf{T}_2$ and $\mathbf{S}_2$ in place of $\mathbf{X}$, $\mathbf{T}$ and $\mathbf{S}$, respectively.

The Bartlett correction factor is very general, and in some cases it is very difficult or even impossible to particularize its formula for specific regression models. For instance, although Cordeiro et al. (1994) found the Bartlett correction factor for the dispersion models, it has no closed-form for the simplex distribution. As can be seen in (8), we have been able to apply the results for our model. For continuous case, i.e., $\pi_\ell = 0, \forall \ell = 1, \ldots, n$, $\varepsilon_{\gamma_{p_3}}$ vanishes, $\mathbf{\Delta}$ is the identity matrix and the term ε_p in (8) coincides with equation (5) from Botter and Cordeiro (1997). Thus, our result generalizes their work.

An alternative strategy for improving LR testing inference is to use the bootstrap procedure. Suppose that $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\pi}}$ are the restricted MLEs of $\boldsymbol{\mu}$, $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$ from the original dataset. In the parametric bootstrap case, B pseudo-samples with size n are generated from ZAGLM($\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\pi}}$), and, for each b, $b = 1, \ldots, B$, the regressor coefficients $\hat{\boldsymbol{\lambda}}^{(b)}$ are estimated and the LR statistic is calculated, as:

$$\text{LR}^{(b)} = 2\left[l\left(\hat{\boldsymbol{\lambda}}^{(b)}\right) - l\left(\tilde{\boldsymbol{\lambda}}^{(b)}\right)\right].$$

The null hypothesis (7) is rejected if $p^\star = (k+1)/(B+1)$ is smaller than or equal to the significance level α. Here, k is the number of bootstrap replications in which $\text{LR}^{(b)}$ is greater than the LR statistic computed using the original sample and $p^\star$ is the bootstrap p-value.

Rocke (1989) proposed bootstrap resampling to estimate the Bartlett correction factor in the following form: B parametric bootstrap samples, imposing the null hypothesis, are produced and the Bartlett-corrected bootstrap test statistic is computed as

$$\text{LR}_{\text{boot2}} = \frac{q\text{LR}}{\overline{\text{LR}^\star}},$$

where $\overline{\text{LR}}^{\star} = B^{-1}\sum_{b=1}^{B} \text{LR}^{(b)}$, the average of all bootstrap statistics. The statistic $\text{LR}_{\text{boot2}} \sim \chi_q^2$, for more details on the Bartlett corrections, included the bootstrap Bartlett adjustment, see Cordeiro and Cribari-Neto (2014).

4 Numerical results

This section presents the results of Monte Carlo simulation studies performed to evaluate the finite sample performance of the following tests: the likelihood ratio statistic (LR), the Bartlett-corrected statistic (LR_c), the bootstrap test based on $p^{\star}$ (LR_{boot1}), and the Bartlett-corrected bootstrap statistic (LR_{boot2}). For each considered scenario, we used 5,000 Monte Carlo replications and $B = 1,000$. The simulations were performed using the Ox language (Doornik 2009).

The following ZAGA regression model was considered in the simulations:

$$\begin{cases} h_1(\mu_l) = \beta_1 + \beta_2 x_{2l} + \beta_3 x_{3l}, \\ h_2(\phi_l) = \delta_1 + \delta_2 t_{2l} + \delta_3 t_{3l}, \\ h_3(\pi_l) = \gamma_1 + \gamma_2 s_{2l} + \gamma_3 s_{3l}, \end{cases} \tag{9}$$

in which we considered $x_{2l} = t_{2l} = s_{2l}$ and $x_{3l} = t_{3l} = s_{3l}$. First, we evaluated the null rejection rates of the different tests considering three nominal levels (1%, 5% and 10%) and three sample sizes (50, 75 and 100). We did not consider sample sizes smaller than 50, because in ZAR models, this leads to bootstrap samples with a very small number of observations equal to zero or a very small number of observations greater than zero. As a consequence, the parameters cannot be estimated in many bootstrap samples when sample size is smaller than 50.

Initially, we considered four scenarios to test the following hypothesis:

$$\begin{cases} \mathcal{H} : \beta_3 = 0, \delta_3 = 0, \gamma_3 = 0, \\ \mathcal{A} : \text{violation of at least one equality.} \end{cases} \tag{10}$$

Note that in (10), we are testing if the distribution of the response variable is a function of the covariate x_3. In Scenarios 1 to 3, the covariate values were taken as random draws of the standard uniform distribution. In Scenario 4, x_2 and x_3 were taken from the Gamma distribution and from inverse Gaussian distribution, respectively, considering for these distributions the same mean and the same variance of the standard uniform distribution. In all scenarios, covariate values were kept fixed in the Monte Carlo replicates.

In Scenario 1, we considered $\beta_1 = 3.0$, $\beta_2 = 2.0$, $\beta_3 = 0.0$, $\delta_1 = 4.0$, $\delta_2 = -2.0$, $\delta_3 = 0.0$, $\gamma_1 = 0.1$, $\gamma_2 = -1.0$ and $\gamma_3 = 0.0$. These parameter values yielded $\mu \in (20.08, 148.42)$, $\phi \in (7.38, 54.60)$, and $\pi \in (0.289, 0.525)$. In the second scenario, we changed the value of γ_1 to 0.8, which yielded $\pi \in (0.450, 0.690)$. In the third scenario, the value of δ_1 was changed to 2.0, which yielded $\phi \in (1.00, 7.39)$. Finally, in Scenario 4, we changed the distribution used to generate the covariates as mentioned before.

Table 2 Null rejection rates (%) of tests for three parameters

Scenario	Test n	$\alpha = 1\%$ 50	75	100	$\alpha = 5\%$ 50	75	100	$\alpha = 10\%$ 50	75	100
1	LR	2.74	2.00	1.50	8.90	7.46	6.16	15.52	14.08	12.18
	LR_c	1.18	0.84	0.92	5.50	5.40	4.86	10.12	10.34	9.44
	LR_{boot1}	0.98	0.84	0.86	5.18	5.22	4.98	9.84	10.62	9.70
	LR_{boot2}	1.06	0.90	0.88	5.34	5.26	5.06	9.96	10.40	9.76
2	LR	3.56	2.32	1.18	10.88	8.58	6.96	18.60	14.36	13.00
	LR_c	0.98	0.92	0.74	5.04	5.22	4.30	10.28	9.94	9.52
	LR_{boot1}	0.72	0.94	0.76	4.14	5.14	4.40	8.66	9.72	9.50
	LR_{boot2}	0.94	1.08	0.82	4.26	5.16	4.54	9.02	9.82	9.70
3	LR	2.42	1.28	1.50	8.12	6.82	6.42	14.82	12.64	12.08
	LR_c	0.98	0.70	0.90	5.14	4.62	5.20	9.68	9.52	9.76
	LR_{boot1}	0.96	0.80	1.08	5.04	4.80	5.30	9.20	9.62	9.88
	LR_{boot2}	0.98	0.76	0.96	5.10	4.86	5.32	9.30	9.70	9.86
4	LR	4.00	1.86	1.52	12.30	7.32	6.56	20.56	13.20	12.62
	LR_c	1.08	1.12	0.92	5.06	4.64	4.64	10.52	9.70	9.74
	LR_{boot1}	0.84	1.06	0.98	4.78	4.74	4.76	9.58	9.78	9.78
	LR_{boot2}	1.04	1.12	1.00	4.94	4.76	4.90	9.70	9.94	9.88

Table 2 presents the null rejection rates for the test (10) in the four scenarios described above. The LR test is considerably liberal for $n = 50$, reaching, for example, a rejection rate of 4.00 when $\alpha = 1\%$ in Scenario 4. Even when $n = 100$, in all scenarios, the rejection rates of the LR test were not close to the nominal levels. On the other hand, in general, the three improved tests presented rejection rates close to the nominal levels even when $n = 50$. For example, in Scenario 4, for $\alpha = 1\%$ and $n = 50$, the rejection rates of LR_c, LR_{boot1} and LR_{boot2} were 1.08, 0.84 and 1.04, respectively. There were some exceptions, mainly in Scenario 2, in which the rejection rates of the test based on the LR_{boot1} statistic were not very close to the nominal levels.

We also performed simulation studies considering a hypothesis with a single parameter. We considered Scenario 1, and three different null hypotheses: $\mathcal{H} : \beta_3 = 0$, $\mathcal{H} : \delta_3 = 0$, and $\mathcal{H} : \gamma_3 = 0$. Table 3 presents the null rejection rates for these tests. The LR test is also considerably liberal for the two tests related with the continuous component of the model, but it performs well for the test related with the discrete component of the model. On the other hand, the improved LR tests have null rejection rates close to the nominal levels for the three hypothesis, even when $n = 50$.

In Tables 4, 5, 6, we present the simulation results to evaluate the non-null rejection rates of the tests (power) based on the statistics LR_c, LR_{boot1} and LR_{boot2}. The LR statistic is not included in this study, because our simulations showed that it is oversized. We considered the following three sets of hypotheses:

- $\mathcal{H} : \beta_3 = 0, \delta_3 = 0, \gamma_3 = 0$ versus $\mathcal{A} : \beta_3 \neq 0, \delta_3 \neq 0, \gamma_3 \neq 0$, considering in $\mathcal{A}$ $\beta_3 = \delta_3 = \gamma_3 = \tau$ (Table 4)

Table 3 Null rejection rates (%) for tests for a single parameter in Scenario 1

$\mathcal{H}$	Test n	$\alpha = 1\%$ 50	75	100	$\alpha = 5\%$ 50	75	100	$\alpha = 10\%$ 50	75	100
$\beta_3 = 0$	LR	2.26	1.96	1.68	8.98	7.18	6.36	14.94	13.26	12.24
	LR_c	0.96	0.86	0.92	4.34	4.94	4.70	9.82	9.42	9.58
	LR_{boot1}	0.96	0.80	1.12	4.62	5.40	4.92	9.94	9.98	9.80
	LR_{boot2}	1.02	0.86	1.12	4.70	5.30	4.90	9.94	10.10	9.92
$\delta_3 = 0$	LR	2.38	1.44	1.66	8.12	6.68	6.00	13.98	12.40	11.34
	LR_c	0.80	0.76	1.06	5.06	4.96	4.78	9.94	9.94	9.56
	LR_{boot1}	0.80	0.84	1.02	4.90	5.10	4.96	9.58	9.90	9.58
	LR_{boot2}	0.88	0.84	1.10	4.86	5.00	5.02	9.60	10.06	9.52
$\gamma_3 = 0$	LR	1.06	1.32	1.14	5.74	6.12	4.98	11.08	11.76	10.10
	LR_c	0.76	1.16	0.90	4.84	5.48	4.58	10.00	10.84	9.74
	LR_{boot1}	0.90	1.18	0.96	4.98	5.52	4.62	10.12	11.12	9.88
	LR_{boot2}	0.90	1.14	1.06	5.00	5.64	4.70	10.18	11.08	9.78

Table 4 Non-null rejection rates (%) for $\mathcal{H} : \beta_3 = 0, \delta_3 = 0, \gamma_3 = 0$ - Scenario 1 with $\beta_3 = \delta_3 = \gamma_3 = \tau$

$n = 50$

τ	$\alpha = 5\%$ LR_c	LR_{boot1}	LR_{boot2}	$\alpha = 10\%$ LR_c	LR_{boot1}	LR_{boot2}
−0.50	79.3	78.3	78.7	87.4	87.1	87.3
−0.25	27.3	27.0	27.4	40.7	39.7	39.8
0.25	30.1	28.9	29.3	43.6	42.4	42.6
0.50	84.2	82.8	83.3	91.5	90.3	90.4

$n = 100$

τ	$\alpha = 5\%$ LR_c	LR_{boot1}	LR_{boot2}	$\alpha = 10\%$ LR_c	LR_{boot1}	LR_{boot2}
−0.50	96.3	96.2	96.4	98.3	98.3	98.3
−0.25	53.8	53.8	54.2	66.7	66.8	66.9
0.25	56.3	56.8	56.8	68.6	69.0	69.0
0.50	98.9	98.8	98.9	99.6	99.6	99.6

- $\mathcal{H} : \beta_3 = 0, \delta_3 = 0, \gamma_3 = 0$ versus $\mathcal{A} : \beta_3 \neq 0, \delta_3 \neq 0, \gamma_3 \neq 0$, considering in $\mathcal{A}$ $\beta_3 = \tau, \delta_3 = \gamma_3 = 0$ (Table 5)
- $\mathcal{H} : \beta_3 = 0$ versus $\mathcal{A} : \beta_3 \neq 0$, considering in $\mathcal{A}$ $\beta_3 = \tau$, (Table 6)

where $\tau = -0.50, -0.25, 0.25$ and 0.50. The remaining parameters were as in Scenario 1.

For the three tables, with fixed values of α, n and τ, the non-null rejection rates were similar in the three tests. As expected, for all cases, the non-null rejection rates increased as n and the absolute value of τ grew.

Table 5 Non-null rejection rates (%) for $\mathcal{H}: \beta_3 = 0, \delta_3 = 0, \gamma_3 = 0$ - Scenario 1 with $\beta_3 = \tau, \delta_3 = \gamma_3 = 0$

$n = 50$						
	$\alpha = 5\%$			$\alpha = 10\%$		
τ	LR_c	LR_{boot1}	LR_{boot2}	LR_c	LR_{boot1}	LR_{boot2}
−0.50	80.2	79.2	79.4	88.8	88.0	88.0
−0.25	28.0	27.0	27.5	40.4	39.6	39.8
0.25	27.4	26.3	26.7	39.6	39.1	39.2
0.50	80.3	79.0	79.6	88.2	87.6	87.8
$n = 100$						
	$\alpha = 5\%$			$\alpha = 10\%$		
τ	LR_c	LR_{boot1}	LR_{boot2}	LR_c	LR_{boot1}	LR_{boot2}
−0.50	97.6	97.6	97.6	98.8	98.8	98.8
−0.25	54.2	54.8	54.5	67.2	67.2	67.5
0.25	46.7	47.0	47.2	59.2	59.4	59.5
0.50	97.9	97.9	98.0	99.1	99.1	99.1

Table 6 Non-null rejection rates (%) for tests for $\mathcal{H}: \beta_3 = 0$ - Scenario 1 with $\beta_3 = \tau$

$n = 50$						
	$\alpha = 5\%$			$\alpha = 10\%$		
τ	LR_c	LR_{boot1}	LR_{boot2}	LR_c	LR_{boot1}	LR_{boot2}
−0.50	88.6	88.5	88.7	93.7	93.6	93.6
−0.25	37.6	37.9	38.3	51.2	51.3	51.4
0.25	37.4	37.7	37.5	50.4	50.4	50.6
0.50	88.3	88.2	88.3	93.6	93.5	93.4
$n = 100$						
	$\alpha = 5\%$			$\alpha = 10\%$		
τ	LR_c	LR_{boot1}	LR_{boot2}	LR_c	LR_{boot1}	LR_{boot2}
−0.50	99.7	99.7	99.7	99.9	99.9	99.9
−0.25	64.4	65.3	65.3	75.9	76.6	76.6
0.25	71.0	72.3	72.4	81.6	82.2	82.1
0.50	99.8	99.8	99.9	99.9	99.9	99.9

5 Application

This section presents an application to real data using the LR test and its improved versions. The data are part of the work presented in Rocha et al. (2009) and refer to the production of mycotoxin FB_2 in corn grains in Brazil. *Fusarium verticillioides* is a species of fungus that commonly produces mycotoxins in corn grains. When these substances are produced in high quantity, the corn grains become improper for consumption.

The dataset consists of 200 unit samples, each composed of 30 gs of corn grains. The response variable is the quantity of mycotoxin FB_2 (measured in μg/g) and the covariates are the percentage of water activity (x_2) and the percentage of grains with *F. verticillioides* infection (x_3). A total of 51 out of 200 unit samples did not contain mycotoxin FB_2.

We fitted ZAGA and ZAIG regression models for the quantity of mycotoxin FB_2 with the following three systematic components:

$$\begin{cases} \log(\mu_l) & = \beta_1 + \beta_2 x_{2l} + \beta_3 x_{3l}, \\ \log(\phi_l) & = \delta_1 + \delta_2 x_{2l} + \delta_3 x_{3l}, \\ \operatorname{logit}(\pi_l) & = \gamma_1 + \gamma_2 x_{2l} + \gamma_3 x_{3l}. \end{cases} \tag{11}$$

Note that we considered in (11) the two covariates in the three components of the model. We used a logarithmic link function in the submodels for μ and ϕ and a logit link in the submodel for π. Diagnostic analysis omitted here for the sake of brevity suggested that the ZAGA regression model is adequate to fit these data and that the ZAIG regression model produces a worse fit. In addition, the AIC value is much lower for the ZAGA regression model (373.09) than for the ZAIG regression model (429.37).

First, we tested if the three parameters related to the percentage of grains with Fusarium verticillioides are all equal to zero. The p-values of the four tests considered here are greater than 0.1. Therefore, there is no evidence that the quantity of mycotoxin FB_2 is a function of the the percentage of grains with *F. verticillioides* and we excluded this covariate from the model.

Second, we fitted a model using water activity as the single covariate and tested if the three parameters related to this covariate were all equal to zero. The second and third columns of Table 7 present the results. Note that the value of the statistic is greater for the LR test than for the corrected test, in agreement with the simulation results. However, all tests yielded the same conclusion, that the quantity of mycotoxin FB_2 is a function of the water activity at the usual nominal levels.

Finally, using the same model, we tested if each of the parameters of the model was equal to zero; the results are presented in the last 6 columns of Table 7. At the usual nominal levels, all statistics also yielded the same conclusion for the three tests. The mean of the continuous component of the quantity of mycotoxin FB_2 and the probability of the quantity of mycotoxin FB_2 assuming a zero value are functions of the water activity, but there is no evidence that the dispersion parameter of the continuous component of the quantity of mycotoxin FB_2 was a function of the water activity.

Table 8 presents the estimates of the parameters and their standard errors for the final model. To interpret the estimates of the parameters, the table also presents the exponential of the estimates. The results indicated that, for every percentage point increase in the water activity, the mean quantity of mycotoxin FB_2, given that there is some FB_2 (mean of the continuous component of the response variable), increased by 5.5%. It was also estimated that, for every percentage point increase in the water activity, the odds of a random sample of 30 gs of corn not containing FB_2, decreased by 12.8%.

Similar to Melo et al. (2022), we randomly selected a subset to illustrate that the conclusions of different tests may be different. We selected the sample using a binomial random variable with probability of success of 0.2 and obtained a dataset with $n = 41$.

Table 7 Test results for the ZAGA model in the complete mycotoxin database

Test	$\mathcal{H}: \beta_2 = \delta_2 = \gamma_2 = 0$		$\mathcal{H}: \beta_2 = 0$		$\mathcal{H}: \delta_2 = 0$		$\mathcal{H}: \gamma_2 = 0$	
	Value	P-value	Value	P-value	Value	P-value	Value	P-value
LR	32.605	< 0.0001	12.951	0.0003	0.893	0.3447	19.159	< 0.0001
LR_c	31.851	< 0.0001	12.423	0.0004	0.862	0.3532	18.838	< 0.0001
LR_{boot1}	–	< 0.0010	–	< 0.0010	–	0.3796	–	< 0.0010
LR_{boot2}	32.348	< 0.0001	11.907	0.0006	0.852	0.3560	18.892	< 0.0001

Table 8 Estimates and standard errors for the final ZAGA model in the completed mycotoxin database

Submodel	Covariate	*Estimate*	St. Error	Exp(estimate)
μ	Intercept	−4.9084	1.1917	–
	Water activity	0.0535	0.0144	1.055
ϕ	Intercept	−0.0350	0.0513	–
π	Intercept	10.0101	2.6123	–
	Water activity	−0.1371	0.0326	0.872

Table 9 Test results for the ZAGA model in the reduced mycotoxin database

	$\mathcal{H}:\beta_2=\delta_2=\gamma_2=0$		$\mathcal{H}:\beta_2=0$		$\mathcal{H}:\delta_2=0$		$\mathcal{H}:\gamma_2=0$	
Test	Value	P-value	Value	P-value	Value	P-value	Value	P-value
LR	12.986	0.0047	4.185	0.0408	3.871	0.0491	5.09	0.0241
LR_c	11.024	0.0116	3.178	0.0746	2.856	0.0910	4.368	0.0366
LR_{boot1}	–	0.0120	–	0.0619	–	0.0679	–	0.0240
LR_{boot2}	11.272	0.0103	3.357	0.0669	3.549	0.0596	5.681	0.0171

Table 9 presents for this reduced dataset the results of the same tests performed before for the complete data. Note that considering a significance level of 1% or 5%, the conclusions were different between the LR tests and the improved LR tests. Based on the results of the simulation studies, if the reduced database were the true one, we would rely on the conclusions reached when using the improved LR tests.

6 Concluding remarks

Response variables that are discrete at zero and continuous on the positive real numbers are common in many areas and they are usually fitted using zero-adjusted generalized linear models. In many regression models, the likelihood ratio test is used to perform hypothesis testing, especially when the null hypothesis involves more than one parameter. However, the likelihood ratio test is considerably liberal (oversized) in the class of ZAGLMs when the sample size is small. In this work, we derived an analytical Bartlett-corrected likelihood ratio test and considered two bootstrap-based corrected likelihood ratio tests. We developed Monte Carlo simulation studies that showed that the null rejection rates of the three improved tests are close to nominal levels for small sample sizes. We also concluded that the three improved likelihood ratio tests considered here have similar power. An application illustrated the usefulness of the improved statistics.

Zero-adjusted regression models are a wide class of regression models that contain the ZAGLMs. There are no previous studies that have evaluated the performance of improved hypothesis tests that simultaneously involve parameters of the continuous and discrete component of the model. Therefore, this work is the first to show that the improved likelihood ratio tests perform well for these kinds of hypotheses, which are useful when one wants to test if the distribution of the response variable is a function of a covariate.

Based on the results of the simulation studies and the features of the three corrected likelihood ratio tests considered here, we suggest that practitioners use the analytical Bartlett-corrected likelihood ratio test when the sample size is small and they want to perform hypothesis testing in the class of ZAGLMs. Our recommendation is based especially on two reasons. First, the performances of the tests related to size and power were similar, but in one of the scenarios considered here, the null rejection rate of the bootstrap corrected test was slightly size distorted. The other reason is that bootstrap uses randomization. For this reason, when the p-value of the test is close to the significance level, two practitioners can reach different conclusion for the same database and hypothesis. This does not happen with the analytical correction.

Appendix

The remaining quantities to define the Bartlett correction factor, see equation (8), are:

$$
\begin{aligned}
\varepsilon_{\beta_{p_1}} &= \frac{1}{4}\mathrm{tr}\left\{\boldsymbol{\Delta\Phi H}_1\boldsymbol{Z}^2_{\beta d}\right\} - \frac{1}{3}\mathbf{1}^\top\boldsymbol{\Delta\Phi G Z}^{(3)}_{\beta}(\boldsymbol{F}+\boldsymbol{G})\boldsymbol{\Phi\Delta}\mathbf{1}\\
&+ \frac{1}{12}\mathbf{1}^\top\boldsymbol{\Delta\Phi F}\left(2\boldsymbol{Z}^{(3)}_{\beta} + 3\boldsymbol{Z}_{\beta d}\boldsymbol{Z}_{\beta}\boldsymbol{Z}_{\beta d}\right)\boldsymbol{F\Phi\Delta}\mathbf{1},\\
\varepsilon_{\delta_{p_2}} &= \frac{1}{4}\mathrm{tr}\left\{\boldsymbol{\Delta H}_2\boldsymbol{Z}^2_{\delta d}\right\} + \frac{1}{12}\mathbf{1}^\top\boldsymbol{\Delta D}_3\boldsymbol{\Phi}_1^3\left(2\boldsymbol{Z}^{(3)}_{\delta} + 3\boldsymbol{Z}_{\delta d}\boldsymbol{Z}_{\delta}\boldsymbol{Z}_{\delta d}\right)\boldsymbol{\Phi}_1^3\boldsymbol{D}_3\boldsymbol{\Delta}\mathbf{1}\\
&+ \frac{1}{4}\mathbf{1}^\top\boldsymbol{\Delta D}_2\boldsymbol{\Phi}_1\boldsymbol{\Phi}_2\left(\boldsymbol{Z}_{\delta d}\boldsymbol{Z}_{\delta}\boldsymbol{Z}_{\delta d} - 2\boldsymbol{Z}^{(3)}_{\delta}\right)\boldsymbol{\Phi}_2\boldsymbol{\Phi}_1\boldsymbol{D}_2\boldsymbol{\Delta}\mathbf{1}\\
&+ \frac{1}{2}\mathbf{1}^\top\boldsymbol{\Delta D}_3\boldsymbol{\Phi}_1^3\boldsymbol{Z}_{\delta d}\boldsymbol{Z}_{\delta}\boldsymbol{Z}_{\delta d}\boldsymbol{\Phi}_2\boldsymbol{\Phi}_1\boldsymbol{D}_2\boldsymbol{\Delta}\mathbf{1},\\
\varepsilon_{\gamma_{p_3}} &= \frac{1}{4}\mathrm{tr}\left\{\boldsymbol{H}_3\boldsymbol{Z}^2_{\gamma d}\right\} - \frac{1}{3}\mathbf{1}^\top\boldsymbol{G}_{\pi}\boldsymbol{Z}^{(3)}_{\gamma}(\boldsymbol{F}_{\pi}+\boldsymbol{G}_{\pi})\mathbf{1}\\
&+ \frac{1}{12}\mathbf{1}^\top\boldsymbol{F}_{\pi}\left(2\boldsymbol{Z}^{(3)}_{\gamma} + 3\boldsymbol{Z}_{\gamma d}\boldsymbol{Z}_{\gamma}\boldsymbol{Z}_{\gamma d}\right)\boldsymbol{F}_{\pi}\mathbf{1},
\end{aligned}
$$

$\boldsymbol{F} = \mathrm{diag}\{f_{11}, \ldots, f_{nn}\}$, $\boldsymbol{G} = \mathrm{diag}\{g_{11}, \ldots, g_{nn}\}$, $\boldsymbol{F}_{\pi} = \mathrm{diag}\{f_{\pi 11}, \ldots, f_{\pi nn}\}$, $\boldsymbol{G}_{\pi} = \mathrm{diag}\{g_{\pi 11}, \ldots, g_{\pi nn}\}$, $\boldsymbol{H}_i = \mathrm{diag}\{h_{i,11}, \ldots, h_{i,nn}\}$, $i = 1, 2, 3$, where

$$
\begin{aligned}
f_{\ell\ell} &= \frac{1}{V_\ell}\frac{d\mu_\ell}{d\eta_{1\ell}}\frac{d^2\mu_\ell}{d\eta_{1\ell}^2}, \quad g_{\ell\ell} = \frac{1}{V_\ell}\frac{d\mu_\ell}{d\eta_{1\ell}}\frac{d^2\mu_\ell}{d\eta_{1\ell}^2} - \frac{1}{V_\ell^2}\frac{dV_\ell}{d\mu_\ell}\left(\frac{d\mu_\ell}{d\eta_{1\ell}}\right)^3,\\
f_{\pi\ell\ell} &= \frac{1}{V_{\pi\ell}}\frac{d\pi_\ell}{d\eta_{3\ell}}\frac{d^2\pi_\ell}{d\eta_{3\ell}^2}, \quad g_{\pi\ell\ell} = \frac{1}{V_{\pi\ell}}\frac{d\pi_\ell}{d\eta_{3\ell}}\frac{d^2\pi_\ell}{d\eta_{3\ell}^2} - \frac{1}{V_{\pi\ell}^2}\frac{dV_{\pi\ell}}{d\pi_\ell}\left(\frac{d\pi_\ell}{d\eta_{3\ell}}\right)^3,\\
h_{1,\ell\ell} &= w_\ell^2\left[\frac{2}{V_\ell}\left(\frac{dV_\ell}{d\mu_\ell}\right)^2 - \frac{d^2V_\ell}{d\mu_\ell^2}\right] + \frac{1}{V_\ell}\frac{d^2\mu_\ell}{d\eta_{1\ell}^2}\left[\frac{d^2\mu_\ell}{d\eta_{1\ell}^2} - 4w_\ell\frac{dV_\ell}{d\mu_\ell}\right],
\end{aligned}
$$

$$h_{2,\ell\ell} = \frac{d^4 d_2(\phi_\ell)}{d\phi_\ell^4}\left(\frac{d\phi_\ell}{d\eta_{2\ell}}\right)^4 + 2\frac{d^3 d_2(\phi_\ell)}{d\phi_\ell^3}\left(\frac{d\phi_\ell}{d\eta_{2\ell}}\right)^2 \frac{d^2\phi_\ell}{d\eta_{2\ell}^2} - \frac{d^2 d_2(\phi_\ell)}{d\phi_\ell^2}\left(\frac{d^2\phi_\ell}{d\eta_{2\ell}^2}\right)^2,$$

$$h_{3,\ell\ell} = 2\left[\frac{1}{\pi_\ell^2(1-\pi_\ell)^2} + \frac{(1-2\pi_\ell)^2}{\pi_\ell^3(1-\pi_\ell)^3}\right]\left(\frac{d\pi_\ell}{d\eta_{3\ell}}\right)^4 - \frac{4(1-2\pi_\ell)}{\pi_\ell^2(1-\pi_\ell)^2}\left(\frac{d\pi_\ell}{d\eta_{3\ell}}\right)^2 \frac{d^2\pi_\ell}{d\eta_{3\ell}^2} + \frac{1}{\pi_\ell(1-\pi_\ell)}\left(\frac{d^2\pi_\ell}{d^2\eta_{3\ell}}\right)^2.$$

Although deriving the expression for ε_p entails a great deal of algebra, this expression only involves simple operations of diagonal matrices, i.e. they are simple expressions to implement.

Acknowledgements We acknowledge the two referees' suggestions, which helped us to improve this work.

Funding This work was partially supported by São Paulo Research Foundation (FAPESP), grant number 2020/16334-9.

References

Araújo MC, Cysneiros AH, Montenegro LC (2020) Improved heteroskedasticity likelihood ratio tests in symmetric nonlinear regression models. Stat Pap 61:167–188

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proceedings of the royal society of London A: Mathematical, Physical and Engineering Sciences 160(901):268–282

Bayer FM, Cribari-Neto F (2013) Bartlett corrections in beta regression models. J Stat Plan Inference 143(3):531–547

Bortoluzzo AB, Claro DP, Caetano MAL, Artes R (2011) Estimating total claim size in the auto insurance industry: a comparison between tweedie and zero-adjusted inverse gaussian distribution. Braz Adm Rev 8(1):37–47

Botter DA, Cordeiro GM (1997) Bartlett corrections for generalized linear models with dispersion covariates. Commun Stat Theory Methods 26(2):279–307

Calsavara VF, Rodrigues AS, Rocha R, Louzada F, Tomazella V, Souza AC, Costa RA, Francisco RPV (2019) Zero-adjusted defective regression models for modeling lifetime data. J Appl Stat 46(13):2434–2459

Cordeiro GM (1983) Improved likelihood ratio statistics for generalized linear models. J R Stat Soc B 45(3):404–413

Cordeiro GM, Cribari-Neto F (2014) An introduction to Bartlett correction and bias reduction. Springer, Berlin

Cordeiro GM, Paula GA, Botter DA (1994) Improved likelihood ratio tests for dispersion models. Int Stat Rev 62(2):257–274

Cox DR, Reid N (1987) Parameter orthogonality and approximate conditional inference. J R S Soc B 49(1):1–39

Das U, Dhar SS, Pradhan V (2018) Corrected likelihood-ratio tests in logistic regression using small-sample data. Commun Stat Theory Methods 47(17):4272–4285

Doornik JA (2009) An object-oriented matrix language: Ox 6. Timberlake Consultants Press, London

Dunn PK, Smyth GK (2018) Generalized linear models with examples in R. Springer, Berlin

Efron B et al (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26

Guedes AC, Cribari-Neto F, Espinheira PL (2020) Modified likelihood ratio tests for unit gamma regressions. J Appl Stat 47(9):1562–1586

Guedes AC, Cribari-Neto F, Espinheira PL (2021) Bartlett-corrected tests for varying precision beta regressions with application to environmental biometrics. PLoS ONE 16(6):e0253349

Hashimoto EM, Ortega EM, Cordeiro GM, Cancho VG, Klauberg C (2019) Zero-spiked regression models generated by gamma random variables with application in the resin oil production. J Stat Comput Simul 89(1):52–70

Heller G, Stasinopoulos M, Rigby B The zero-adjusted inverse gaussian distribution as a model for insurance claims, in: J. Hinde, J. Einbeck, J. Newell (Eds.), Proceedings of the 21th International Workshop on Statistical Modelling, Ireland, Galway, 2006, pp. 226–233

Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14

Lawley DN (1956) A general method for approximating to the distribution of likelihood ratio criteria. Biometrika 43(3/4):295–303

Loose LH, Bayer FM, Pereira TL (2017) Bootstrap bartlett correction in inflated beta regression. Commun Stat-Simul Comput 46(4):2865–2879

Loose LH, Valença DM, Bayer FM (2018) On bootstrap testing inference in cure rate models. J Stat Comput Simul 88(17):3437–3454

Magalhães TM, Gallardo DI (2020) Bartlett and bartlett-type corrections for censored data from a weibull distribution. Stat Op Res Trans 44(1):127–140

Melo TF, Ferrari SL, Cribari-Neto F (2009) Improved testing inference in mixed linear models. Comput Stat Data Anal 53(7):2573–2582

Melo TF, Vargas TM, Lemonte AJ, Moreno-Arenas G (2022) Higher-order asymptotic refinements in the multivariate dirichlet regression model. Commun Stat-Simul Comput 51(1):53–71

Michaelis P, Klein N, Kneib T (2020) Mixed discrete-continuous regression-a novel approach based on weight functions. Stat 9(1):e277

Moulton LH, Weissfeld LA, Laurent RTS (1993) Bartlett correction factors in logistic regression models. Comput Stat Data Anal 15(1):1–11

Nogarotto DC, Azevedo CLN, Bazán JL et al (2020) Bayesian modeling and prior sensitivity analysis for zero-one augmented beta regression models with an application to psychometric data. Braz J Probab Stat 34(2):304–322

Ospina R, Ferrari S (2012) A general class of zero-or-one inflated beta regression models. Comput Stat Data Anal 56(6):1609–1623

Pereira TL, Cribari-Neto F (2014) Modified likelihood ratio statistics for inflated beta regressions. J Stat Comput Simul 84(5):982–998

Pereira GH, Scudilio J, Santos-Neto M, Botter DA, Sandoval MC (2020) A class of residuals for outlier identification in zero adjusted regression models. J Appl Stat 47(10):1833–1847

Rauber C, Cribari-Neto F, Bayer FM (2020) Improved testing inferences for beta regressions with parametric mean link function. Adv Stat Anal 104(4):687–717

Rocha LO, Nakai VK, Braghini R, Reis TA, Kobashigawa E, Corrêa B (2009) Mycoflora and co-occurrence of fumonisins and aflatoxins in freshly harvested corn in different regions of brazil. Int J Mol Sci 10(11):5090–5103

Rocha LO, Reis GM, Fontes LC, Piacentini KC, Barroso VM, Reis TA, Pereira AA, Corrêa B (2017) Association between fum expression and fumonisin contamination in maize from silking to harvest. Crop Prot 94:77–82

Rocke DM (1989) Bootstrap bartlett adjustment in seemingly unrelated regression. J Am Stat Assoc 84(406):598–601

Rubec PJ, Kiltie R, Leone E, Flamm RO, McEachron L, Santi C (2016) Using delta-generalized additive models to predict spatial distributions and population abundance of juvenile pink shrimp in tampa bay, florida. Marine Coastal Fish 8(1):232–243

Sen PN, Singer JM, Lima ACP (2010) From finite sample to asymptotic methods in statistics. Cambridge University Press, Cambridge

Silva AR, Azevedo CL, Bazán JL, Nobre JS (2021) Augmented-limited regression models with an application to the study of the risk perceived using continuous scales. J Appl Stat 48(11):1998–2021

Skovgaard IM (2001) Likelihood asymptotics. Scand J Stat 28(1):3–32

Thomson FJ, Letten AD, Tamme R, Edwards W, Moles AT (2018) Can dispersal investment explain why tall plant species achieve longer dispersal distances than short plant species? New Phytol 217(1):407–415

Tomazella V, Nobre JS, Pereira GH, Santos-Neto M (2019) Zero-adjusted birnbaum-saunders regression model. Statist Probab Lett 149:142–145

Tong E, Mues C, Thomas L (2013) A zero-adjusted gamma model for mortgage loan loss given default. Int J Forecast 29:548–562

Tong EN, Mues C, Brown I, Thomas LC (2016) Exposure at default models with and without the credit conversion factor. Eur J Op Res 252(3):910–920

Ye T, Lachos VH, Wang X, Dey DK (2021) Comparisons of zero-augmented continuous regression models from a Bayesian perspective. Stat Med 40(5):1073–1100

Zamani H, Bazrafshan O (2020) Modeling monthly rainfall data using zero-adjusted models in the semi-arid, arid and extra-arid regions. Meteorol Atmos Phys 132(2):239–253