Electronic Journal of Statistics
Vol. 8 (2014) 1641–1660
ISSN: 1935-7524
DOI: 10.1214/14-EJS936

Estimating equations and diagnostic techniques applied to zero-inflated models for panel data

Maria Kelly Venezuela[*,†] **and Rinaldo Artes**

Insper Institute of Education and Research
Rua Quatá, 300 – Vila Olímpia – São Paulo/SP – Brasil, CEP: 04546-042
e-mail: MariaKV@insper.edu.br*;* RinaldoA@insper.edu.br

Abstract: Many practical studies analyze semi-continuous longitudinal data using, for example, ZAIG (Zero-Adjusted Inverse Gaussian) or BEZI (Beta Zero-Inflated) models as marginal distributions. We develop herein estimating equations analogous to Liang and Zeger's estimating equation of independence and related diagnostic techniques for regression models, with zero-inflated response variable and panel data structure. A simulation study to evaluate some properties of the estimators obtained from the estimating equations and two applications with real data are presented.

MSC 2010 subject classifications: Primary 62J20; secondary 62J12, 62H12.
Keywords and phrases: Estimating equations, diagnostic techniques, zero inflated data.

Received August 2013.

Contents

[*]Corresponding author.
[†]The authors thank the editor and the reviewers for their comments and suggestions.

1. Introduction

Semi-continuous response variables appear in many practical situations. In finance, for instance, losses in loans (zero loss means no default and a positive loss means no payment); in medicine when it is necessary to measure the concentration of a certain substance that may or may not be present in blood. In these cases, the response variable y can be described as

$$y = \begin{cases} 0, & \text{if } B = 1, \\ C, & \text{if } B = 0, \end{cases}$$

where C is a positive random variable (in this paper, continuous) and B is a Bernoulli variable with parameter ν. The probability distributions attributed to model y are called *zero-inflated*.

Some particularly interesting distributions are the Zero-Adjusted Inverse Gaussian distribution, named as ZAIG distribution, e.g. [5, 7], and the Beta Zero-Inflated distribution, named as BEZI distribution [14]. The ZAIG is a zero-inflated distribution, where C follows an inverse normal distribution. In the case of BEZI, the random variable C follows a beta distribution. Ridout et al. [17] describe other inflated distributions for counting data.

In cross section studies, estimates of the regression model parameters for ZAIG or BEZI responses may be obtained by the library GAMLSS [19] for R[1].

The authors in [4] developed generalized estimating equations for zero inflated random variables. In their proposal, there was a regression model for the probability of zero and for the mean of the continuous part of the distribution. The method assumes that C belongs to the class of exponential dispersion models (see [8], for instance) and when it follows a log-normal distribution. Dobbie and Welsh [2] developed estimating equations for zero-inflated counting data.

Other advances in the study of regression models for zero-inflated distributions, in the presence of dependence among the observations and when C is a discrete variable, may be found in [13], who developed random effect models for zero-inflated counting data. Multivariate distributions with zero-inflated marginal distributions may be found in [3].

In this paper we developed diagnostic techniques for regression models, with zero-inflated response variable and panel data structure, applying estimating equations. We consider cases with homogeneous and heterogeneous dispersion parameters. These techniques are based on [20] who considered estimating equations for longitudinal continuous data with probability distribution in the exponential family and under a homogeneous dispersion parameter.

This paper is organized as follows. The next section presents basic concepts of estimating functions. In Section 2 we propose a general model for zero-inflated panel data analysis, analogous to Liang and Zeger's independent estimating equations, whose application to ZAIG and BEZI response variable is presented in Sections 2.2 and 2.3, respectively. Section 2.1 describes an interactive method for estimating parameters. In Section 3 we propose some diagnostics measures.

[1] See [16].

Some properties of the estimators obtained by the proposed estimating equations are evaluated by a simulation study in Section 4. Finally, we analyze a data set with the proposed methodology and then we present our concluding remarks.

2. General model

By definition, any measurable function $\boldsymbol{\psi}$ of the data and of the parameters of interest ($\boldsymbol{\theta}$) is an estimating function. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})^\top$, $i = 1, \dots, n$, be a sample of independent random vectors[2]. Assume that $\boldsymbol{\psi}_i = \boldsymbol{\psi}_i(\mathbf{y}_i; \boldsymbol{\theta})$, $i = 1, \dots, n$, are estimating functions. The concept of estimating function may be extended to the sample by $\boldsymbol{\Psi}(\mathbf{y}; \boldsymbol{\theta}) = \boldsymbol{\Psi}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{\psi}_i(\mathbf{y}_i; \boldsymbol{\theta})$, where $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$ is an $(N \times 1)$ dimensional vector, $N = nT$.

Under general regularity conditions, e.g. [18], we may prove that the estimator $\hat{\boldsymbol{\theta}}$, obtained from $\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}) = 0$ is consistent and that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with null mean vector and covariance matrix given by the inverse of

$$\mathbf{J}(\boldsymbol{\theta}) = \lim_{n \to \infty} n \mathbf{S}_n^\top(\boldsymbol{\theta}) \mathbf{K}_n^{-1}(\boldsymbol{\theta}) \mathbf{S}_n(\boldsymbol{\theta}), \tag{1}$$

where $\mathbf{S}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathrm{E}(\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\theta}^\top}(\mathbf{y}_i; \boldsymbol{\theta}))$ is the *sensibility matrix* and $\mathbf{K}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathrm{E}(\boldsymbol{\Psi}(\mathbf{y}_i; \boldsymbol{\theta}) \boldsymbol{\Psi}^\top(\mathbf{y}_i; \boldsymbol{\theta}))$ is the *variability matrix*.

In the presence of zero-inflated data, let $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})^\top$ be a sample of independent random vectors of the ith experimental unit, $i = 1, \dots, n$, with

$$y_{it} = \begin{cases} 0, & \text{with probability } \nu_{it}, \\ C_{it}, & \text{with probability } (1 - \nu_{it}), \end{cases}$$

where C_{it} is a continuous random variable with a regular probability density function given by $f_{it} = f(y_{it}; \mu_{it}, \sigma_{it})$, with a position parameter μ_{it} and a second parameter given by σ_{it} (by convenience, it will be called a dispersion parameter, although it may indicate any other distribution characteristic).

Consider the existence of three column vectors of fixed covariates $\mathbf{x}_{it}$, $\mathbf{d}_{it}$ and $\mathbf{q}_{it}$ of dimensions p, q and r, respectively, such as

$$g_1(\nu_{it}) = \mathbf{x}_{it}^\top \boldsymbol{\beta}, \qquad g_2(\mu_{it}) = \mathbf{d}_{it}^\top \boldsymbol{\gamma} \qquad \text{and} \qquad g_3(\sigma_{it}) = \mathbf{q}_{it}^\top \boldsymbol{\delta},$$

where $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are parametric vectors and g_1, g_2 and g_3 are continuous, invertible and double differentiated link functions, with $i = 1, \dots, n$ and $t = 1, \dots, T$.

The probability density function of y_{it} is given by

$$p_{it} = p(y_{it}; \nu_{it}, \mu_{it}, \sigma_{it}) \quad = \quad \left\{ \nu_{it}^{\mathbf{I}_{\{y_{it}=0\}}} (1 - \nu_{it})^{\mathbf{I}_{\{y_{it} \neq 0\}}} \right\} \left\{ f_{it}^{\mathbf{I}_{\{y_{it} \neq 0\}}} \right\},$$

where $\mathbf{I}_{y \in A}$ is an indicator variable that assumes the value 1 if y belongs to the set A. The right side of equation p_{it} may be factorized in one term that just

[2] The number of observations T does not need to be necessarily equal among experimental units and can be treated as T_i.

depends on ν_{it} and another one that depends on μ_{it} and σ_{it}, which uses the continuous part of the dependent variable (see [11]).

We propose the use of an estimating function for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)^\top$ that is identical to the score function obtained in the case of independence among y_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, T$ (full independence). In practical terms, the point estimators of the parameters are identical to those obtained under full independence, but the standard errors must be corrected by equation (1) for the presence of dependence among observations of the same sample unit. This result is similar to the one obtained by [10] for independence estimating equations.

Under full independence, the likelihood function for $\boldsymbol{\theta}$ would be given by

$$L(\boldsymbol{\theta}) = L_1(\boldsymbol{\beta})L_2(\boldsymbol{\gamma}, \boldsymbol{\delta}),$$

where

$$\begin{aligned} L_1(\boldsymbol{\beta}) &= \prod_{i=1}^{n}\prod_{t=1}^{T} \nu_{it}^{\mathbf{I}_{\{y_{it}=0\}}}(1-\nu_{it})^{\mathbf{I}_{\{y_{it}\neq 0\}}}, \\ L_2(\boldsymbol{\gamma}, \boldsymbol{\delta}) &= \prod_{i,t:y_{it}\neq 0} f_{it}^{\mathbf{I}_{\{y_{it}\neq 0\}}}. \end{aligned}$$

Consequently, the logarithm of the likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)^\top$ would be given by

$$\ell(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\beta}) + \ell_2(\boldsymbol{\gamma}, \boldsymbol{\delta}), \tag{2}$$

where

$$\begin{aligned} \ell_1(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{t=1}^{T} \ell_{it}(\nu_{it}) = \sum_{i=1}^{n}\sum_{t=1}^{T} \left[\mathbf{I}_{\{y_{it}=0\}}\ln(\nu_{it}) + \mathbf{I}_{\{y_{it}\neq 0\}}\ln(1-\nu_{it})\right], \\ \ell_2(\boldsymbol{\gamma}, \boldsymbol{\delta}) &= \sum_{i,t:y_{it}\neq 0} \ell_{it}(\mu_{it}, \sigma_{it}) = \sum_{i,t:y_{it}\neq 0} \ln(f_{it}), \end{aligned}$$

From (2) the estimating equation, which coincides with the score function obtained in case of full independence, is given by

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{M}_i^\top \boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i^{-1}\mathbf{a}_i, \tag{3}$$

where $\mathbf{M}_i = \mathrm{diag}\{\mathbf{X}_i, \mathbf{D}_i, \mathbf{Q}_i\}$, $\boldsymbol{\Lambda}_i = \mathrm{diag}\{\mathbf{G}_{1i}, \mathbf{G}_{2i}, \mathbf{G}_{3i}\}$, $\boldsymbol{\Omega}_i = \mathrm{diag}\{\mathbf{V}_{1i}^{-1}, \mathbf{I}_T, \mathbf{I}_T\}$ and $\mathbf{a}_i = ((\jmath_i - \boldsymbol{\nu}_i)^\top, \mathbf{u}_i^\top, \mathbf{m}_i^\top)^\top$, with $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})^\top$, $\mathbf{D}_i = (\mathbf{d}_{i1}, \ldots, \mathbf{d}_{iT})^\top$, $\mathbf{Q}_i = (\mathbf{q}_{i1}, \ldots, \mathbf{q}_{iT})^\top$, $\mathbf{G}_{1i} = \mathrm{diag}\{\frac{\partial \nu_{it}}{\partial g_1(\nu_{it})}\}$, $\mathbf{G}_{2i} = \mathrm{diag}\{\frac{\partial \mu_{it}}{\partial g_2(\mu_{it})}\mathbf{I}_{\{y_{it}\neq 0\}}\}$, $\mathbf{G}_{3i} = \mathrm{diag}\{\frac{\partial \sigma_{it}}{\partial g_3(\sigma_{it})}\}$, $\mathbf{V}_{1i}^{-1} = \mathrm{diag}\{\nu_{it}(1-\nu_{it})\}$, $\jmath_i = (\mathbf{I}_{\{y_{i1}=0\}}, \ldots, \mathbf{I}_{\{y_{iT}=0\}})^\top$, $\boldsymbol{\nu}_i = (\nu_{i1}, \ldots, \nu_{iT})^\top$, $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})^\top$ and $\mathbf{m}_i = (m_{i1}, \ldots, m_{iT})^\top$. Besides,

$$u_{it} = \begin{cases} 0, & \text{if } y_{it} = 0 \\ \frac{\partial \ell_{it}(\mu_{it}, \sigma)}{\partial \mu_{it}}, & \text{if } y_{it} \neq 0 \end{cases}, \qquad m_{it} = \begin{cases} 0, & \text{if } y_{it} = 0 \\ \frac{\partial \ell_{it}(\mu_{it}, \sigma_{it})}{\partial \sigma_{it}}, & \text{if } y_{it} \neq 0 \end{cases},$$

$i = 1, \ldots, n$ and $t = 1, \ldots, T$.

Assuming the existence of an structure of dependence among the components of $\mathbf{y}_i$, from (1), we have $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\xrightarrow{\mathcal{D}}N_p(\mathbf{0},\mathbf{J}^{-1})$, with $\mathbf{S}_n(\boldsymbol{\theta})=-\sum_{i=1}^n \mathbf{M}_i^\top \mathbf{W}_i\mathbf{M}_i$ and $\mathbf{K}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{A}_i\mathrm{Cov}(\mathbf{a}_i)\mathbf{A}_i^\top$, with $\mathbf{A}_i = \mathbf{M}_i^\top\boldsymbol{\Lambda}_i\boldsymbol{\Omega}_i^{-1}$, $\mathbf{W}_i = \boldsymbol{\Lambda}_i\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i\boldsymbol{\Lambda}_i$ and

$$\mathbf{B}_i = \begin{pmatrix} \mathbf{I}_T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\Delta\{\mathrm{E}(\dot{\mathbf{u}}_i)\} & -\Delta\{\mathrm{E}(\dot{\mathbf{u}}_{i\sigma})\} \\ \mathbf{0} & -\Delta\{\mathrm{E}(\dot{\mathbf{u}}_{i\sigma})\} & -\Delta\{\mathrm{E}(\dot{\mathbf{m}}_i)\} \end{pmatrix},$$

where $\boldsymbol{\Delta}(\cdot)$ is an operator that, when applied to a vector, creates a diagonal matrix with the vector elements at the main diagonal. Besides, $\dot{\mathbf{u}}_i$, $\dot{\mathbf{m}}_i$ and $\dot{\mathbf{u}}_{i\sigma}$ are T-dimensional vectors whose components are given, respectively, by

$$\dot{u}_{it} = \begin{cases} 0, & \text{if } y_{it}=0 \\ \frac{\partial u_{it}}{\partial \mu_{it}}, & \text{if } y_{it}\neq 0 \end{cases}, \qquad \dot{m}_{it} = \begin{cases} 0, & \text{if } y_{it}=0 \\ \frac{\partial m_{it}}{\partial \sigma_{it}}, & \text{if } y_{it}\neq 0 \end{cases} \quad \text{and}$$

$$\dot{u}_{it\sigma} = \begin{cases} 0, & \text{if } y_{it}=0 \\ \frac{\partial u_{it}}{\partial \sigma_{it}}, & \text{if } y_{it}\neq 0 \end{cases}.$$

Moreover, $\mathrm{E}\left(\dot{u}_{it\sigma}\right)$ may also be written as

$$\begin{aligned} \mathrm{E}(\dot{u}_{it}) &= \mathrm{E}(\mathrm{E}(\dot{u}_{it}|y_{it})) = \mathrm{E}\left(0|y_{it}=0\right)\nu_{it} + \mathrm{E}\left(\frac{\partial u_{it}}{\partial \mu_{it}}|y_{it}\neq 0\right)(1-\nu_{it}) \\ &= \mathrm{E}\left(\frac{\partial u_{it}}{\partial \mu_{it}}|y_{it}\neq 0\right)(1-\nu_{it}) \end{aligned}$$

and $\mathrm{E}\left(\dot{m}_{it}\right)$ is defined by the same following way.

2.1. Estimation of the parameters

We may obtain an estimate of $\hat{\boldsymbol{\theta}}$ by the following iterative process

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} - \mathbf{S}_n^{-1}(\hat{\boldsymbol{\theta}}^{(m)})\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}^{(m)}), \tag{4}$$

where $m=0,1,2,\ldots$ indicates the step of the iterative process.

Equation (4) may be presented as reweighted least square iterative process with weight matrix $\mathbf{W}_i$ and a modified dependent variable $\mathbf{z}_i$ given by

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \Big(\sum_{i=1}^n \mathbf{M}_i^\top\hat{\mathbf{W}}_i^{(m)}\mathbf{M}_i\Big)^{-1}\sum_{i=1}^n \mathbf{M}_i^\top\hat{\mathbf{W}}_i^{(m)}\mathbf{z}_i^{(m)}, \tag{5}$$

where $\mathbf{z}_i = \hat{\boldsymbol{\tau}}_i + (\hat{\mathbf{B}}_i\hat{\boldsymbol{\Lambda}}_i)^{-1}\hat{\mathbf{a}}_i$ and $\hat{\boldsymbol{\tau}}_i = \mathbf{M}_i\hat{\boldsymbol{\theta}}$.

The matrix $\mathbf{J}$ may be estimated by

$$\hat{\mathbf{J}} = n\mathbf{S}_n^\top(\hat{\boldsymbol{\theta}})\left[\sum_{i=1}^n \mathbf{A}_i(\hat{\boldsymbol{\theta}})\mathbf{a}_i\mathbf{a}_i^\top\mathbf{A}_i(\hat{\boldsymbol{\theta}})^\top\right]^{-1}\mathbf{S}_n(\hat{\boldsymbol{\theta}}).$$

2.2. ZAIG response variable

Herein, we will apply the results of this section to the case of the ZAIG response variable. From (2), the log-likelihood in case of full independence would be given by

$$\begin{aligned}\ell_1(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathbf{I}_{\{y_{it}=0\}}\ln(\nu_{it}) + \mathbf{I}_{\{y_{it}>0\}}\ln(1-\nu_{it})\right],\\ \ell_2(\boldsymbol{\gamma},\boldsymbol{\delta}) &= \sum_{i,t:y_{it}>0}\left\{-\ln\left(\sqrt{2\pi y_{it}^3}\right) - \ln(\sigma_{it}) - \frac{(y_{it}-\mu_{it})^2}{2y_{it}\mu_{it}^2\sigma_{it}^2}\right\}.\end{aligned}$$

When $y_{it} > 0$, we have

$$u_{it} = \frac{\partial \ell_{it}(\mu_{it},\sigma_{it})}{\partial \mu_{it}} = \frac{1}{\sigma_{it}^2}\frac{1}{\mu_{it}^3}(y_{it}-\mu_{it}) \tag{6}$$

and

$$m_{it} = \frac{\partial \ell_{it}(\mu_{it},\sigma_{it})}{\partial \sigma_{it}} = \frac{1}{\sigma_{it}^3}(s_{it}-\sigma_{it}^2), \tag{7}$$

with $s_{it} = \frac{(y_{it}-\mu_{it})^2}{y_{it}\mu_{it}^2}$.

From (6) and (7), the estimating function (3) is given by

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[\begin{array}{c}\mathbf{X}_i^\top \mathbf{G}_{1i}\mathbf{V}_{1i}(\jmath_i - \boldsymbol{\nu}_i)\\ \mathbf{D}_i^\top \mathbf{G}_{2i}\mathbf{V}_{2i}(\mathbf{y}_i - \boldsymbol{\mu}_i)\\ \mathbf{Q}_i^\top \mathbf{G}_{3i}\mathbf{V}_{3i}(\mathbf{s}_i - \boldsymbol{\sigma}_i)\end{array}\right],$$

where $\mathbf{V}_{2i} = \mathrm{diag}\{(\sigma_{it}^*)^{-2}(\mu_{it}^*)^{-3}\}$, $\mathbf{V}_{3i} = \mathrm{diag}\{(\sigma_{it}^*)^{-3}\}$ and $\boldsymbol{\sigma}_i = ((\sigma_{i1}^*)^2,\ldots,(\sigma_{iT}^*)^2)^\top$, with

$$\sigma_{it}^* = \left\{\begin{array}{rl}0, & \text{if } y_{it}=0\\ \sigma_{it}, & \text{if } y_{it}>0\end{array}\right., \qquad \mu_{it}^* = \left\{\begin{array}{rl}0, & \text{if } y_{it}=0\\ \mu_{it}, & \text{if } y_{it}>0\end{array}\right.,$$

$i = 1,\ldots,n$ and $t = 1,\ldots,T$.

For a ZAIG response, under heterogeneity of the dispersion parameter, we have $\mathbf{a}_i = ((\jmath_i - \boldsymbol{\nu}_i)^\top, (\mathbf{V}_{2i}(\mathbf{y}_i - \boldsymbol{\mu}_i))^\top, (\mathbf{V}_{3i}(\mathbf{s}_i - \boldsymbol{\sigma}_i))^\top)^\top$ and

$$\mathbf{B}_i = \begin{pmatrix}\mathbf{I}_T & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}_{2i}\mathbf{N}_i & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{4i}\mathbf{N}_i\end{pmatrix},$$

where $\mathbf{V}_{4i} = \mathrm{diag}\{2(\sigma_{it}^*)^{-2}\}$.

2.3. BEZI response variable

Herein, we analyze a BEZI response variable case. The corresponding log-likelihood function as described in equation (2), under independence of all observations, should be expressed by

$$
\begin{aligned}
\ell_1(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathbf{I}_{\{y_{it}=0\}}\ln(\nu_{it}) + \mathbf{I}_{\{y_{it}\in(0,1)\}}\ln(1-\nu_{it})\right],\\
\ell_2(\boldsymbol{\gamma},\boldsymbol{\delta}) &= \sum_{i,t:y_{it}\in(0,1)} [\log\Gamma(\sigma_{it}) - \log\Gamma(\mu_{it}\sigma_{it}) - \log\Gamma((1-\mu_{it})\sigma_{it}) +\\
&\quad + (\mu_{it}\sigma_{it}-1)\log y_{it} + ((1-\mu_{it})\sigma_{it}-1)\log(1-y_{it})].
\end{aligned}
$$

When $y_{it} \in (0,1)$, one can say

$$
u_{it} = \frac{\partial \ell_{it}(\mu_{it},\sigma_{it})}{\partial \mu_{it}} = \sigma_{it}(y^*_{it} - \tilde{\mu}^*_{it}) \tag{8}
$$

and

$$
\begin{aligned}
m_{it} &= \frac{\partial \ell_{it}(\mu_{it},\sigma_{it})}{\partial \sigma_{it}} =\\
&= \mu_{it}(y^*_{it} - \tilde{\mu}^*_{it}) + \psi(\sigma_{it}) + \log(1-y_{it}) - \psi((1-\mu_{it})\sigma_{it}),
\end{aligned} \tag{9}
$$

where $\psi(\cdot)$ is a digamma function,

$$
y^*_{it} = \begin{cases} 0, & \text{if } y_{it}=0\\ \log\left(\frac{y_{it}}{1-y_{it}}\right), & \text{if } y_{it}\in(0,1)\end{cases} \quad \text{and}
$$

$$
\tilde{\mu}^*_{it} = \begin{cases} 0, & \text{if } y_{it}=0\\ \psi(\mu_{it}\sigma_{it}) - \psi((1-\mu_{it})\sigma_{it}), & \text{if } y_{it}\in(0,1)\end{cases}.
$$

From (8) and (9), the estimating function (3) is given by

$$
\boldsymbol{\Psi}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\begin{bmatrix} \mathbf{X}_i^\top \mathbf{G}_{1i}\mathbf{V}_{1i}(\jmath_i - \boldsymbol{\nu}_i)\\ \mathbf{D}_i^\top \mathbf{G}_{2i}\mathbf{V}_{2i}(\mathbf{y}^*_i - \tilde{\boldsymbol{\mu}}^*_i)\\ \mathbf{Q}_i^\top \mathbf{G}_{3i}\mathbf{m}_i\end{bmatrix},
$$

where $\mathbf{V}_{2i} = \mathrm{diag}\{\tilde{\sigma}^*_{it}\}$, $\mathbf{y}^*_i = (y^*_{i1},\ldots,y^*_{iT})^\top$, $\tilde{\boldsymbol{\mu}}^*_i = (\tilde{\mu}^*_{i1},\ldots,\tilde{\mu}^*_{iT})^\top$, with

$$
\tilde{\sigma}^*_{it} = \begin{cases} 0, & \text{if } y_{it}=0\\ \sigma_{it}, & \text{if } y_{it}\in(0,1)\end{cases},
$$

$i=1,\ldots,n$ and $t=1,\ldots,T$.

According to (3), the estimating function for BEZI response variable models, under heterogeneity of the dispersion parameter is given by

$$
\mathbf{a}_i = \left((\jmath_i - \boldsymbol{\nu}_i)^\top, (\mathbf{V}_{2i}(\mathbf{y}^*_i - \tilde{\boldsymbol{\mu}}^*_i))^\top, \mathbf{m}_i)^\top\right)^\top.
$$

Matrix $\mathbf{B}_i$ is

$$\mathbf{B}_i = \begin{pmatrix} \mathbf{I}_T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{6i}\mathbf{N}_i & \mathbf{V}_{5i}\mathbf{N}_i \\ \mathbf{0} & \mathbf{V}_{5i}\mathbf{N}_i & \mathbf{V}_{4i}\mathbf{N}_i \end{pmatrix},$$

where

$$\begin{aligned} \mathbf{N}_i &= \operatorname{diag}\left\{(1-\nu_{it})\right\}, \\ \mathbf{V}_{4i} &= \operatorname{diag}\left\{\mu_{it}^2\psi'(\mu_{it}\sigma_{it}) + (1-\mu_{it})^2\psi'((1-\mu_{it})\sigma_{it}) - \psi'(\sigma_{it})\right\}, \\ \mathbf{V}_{5i} &= \operatorname{diag}\left\{\sigma_{it}[\mu_{it}\psi'(\mu_{it}\sigma_{it}) - (1-\mu_{it})\psi'((1-\mu_{it})\sigma_{it})]\right\} \quad \text{and} \\ \mathbf{V}_{6i} &= \operatorname{diag}\left\{\sigma_{it}^2[\psi'(\mu_{it}\sigma_{it}) + \psi'((1-\mu_{it})\sigma_{it})]\right\}, \end{aligned}$$

and $\psi'(\cdot)$ is a trigamma function.

3. Diagnostic techniques

Based on [20], this section presents some diagnostic measures to detect leverage points, influential points and outliers for the models proposed in Section 2.

Equation (5) may be written as

$$\hat{\boldsymbol{\theta}} \;=\; \left(\mathbf{M}^\top\hat{\mathbf{W}}\mathbf{M}\right)^{-1}\mathbf{M}^\top\hat{\mathbf{W}}\mathbf{z}, \tag{10}$$

where $\mathbf{M} = (\mathbf{M}_1^\top, \ldots, \mathbf{M}_n^\top)^\top$, $\hat{\mathbf{W}} = \operatorname{diag}(\hat{\mathbf{W}}_i)$ and $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$. In this case, $\mathbf{M}$ represents a design matrix and $\hat{\mathbf{W}}$ and $\mathbf{z}$ assume the role of a weight matrix and a dependent variable, respectively.

From (10), where $\hat{\mathbf{W}}^{1/2}\mathbf{z}$ may be seen as a response vector [15], the ordinary residual is given by

$$\mathbf{r}_i = \hat{\mathbf{W}}_i^{1/2}(\mathbf{z}_i - \hat{\boldsymbol{\tau}}_i) = (\mathbf{I}_T - \mathbf{H}_i)\hat{\mathbf{W}}_i^{1/2}\mathbf{z}_i, \tag{11}$$

where $\mathbf{H}_i = \hat{\mathbf{W}}_i^{1/2}\mathbf{M}_i(\mathbf{M}^\top\hat{\mathbf{W}}\mathbf{M})^{-1}\mathbf{M}_i^\top\hat{\mathbf{W}}_i^{1/2}$ and $\mathbf{I}_T$ is a T-dimensional identity matrix, with $i = 1, \ldots, n$.

Matrix $\mathbf{H}_i$ has characteristics similar to those of hat matrix from a linear model. Therefore, the main diagonal elements may be used to detect the presence of leverage points [20, 15].

To identify outlier observations, a new residual is defined by standardizing the ordinary residual described in (11). Under the model described in (3), one may consider that $\operatorname{Cov}(\mathbf{r}_i|\boldsymbol{\theta}) \cong \mathbf{C}_i(\boldsymbol{\theta}) = (\mathbf{I}_T - \mathbf{H}_i)\mathbf{W}_i^{1/2}\boldsymbol{\Lambda}_i^{-1}\mathbf{B}_i^{-1}\mathbf{a}_i\mathbf{a}_i^\top\mathbf{B}_i^{-1} \times \boldsymbol{\Lambda}_i^{-1}\mathbf{W}_i^{1/2}(\mathbf{I}_T - \mathbf{H}_i)$. Thus, the standardized residual for observation y_{it} is given by

$$\mathrm{r}_{SDit} = \frac{\mathbf{e}_{(it)}^\top\hat{\mathbf{W}}_i^{1/2}(\mathbf{z}_i - \hat{\boldsymbol{\tau}}_i)}{\sqrt{\mathrm{c}_{it}}} = \frac{\mathbf{e}_{(it)}^\top\hat{\mathbf{W}}_i^{1/2}(\hat{\mathbf{B}}_i\hat{\boldsymbol{\Lambda}}_i)^{-1}\hat{\mathbf{a}}_i}{\sqrt{\mathrm{c}_{it}}}, \tag{12}$$

where $\mathbf{e}_{(it)}$ is a T-dimensional vector assuming the value 1 in the position related to y_{it} and 0 otherwise, and c_{it} is the t-th element of the main diagonal of $\mathbf{C}_i(\hat{\boldsymbol{\theta}})$, with $i = 1, \ldots, n$ and $t = 1, \ldots, T$. This result is not a simple extension of [20].

For repeated measures regression models [20], the Cook distance used to detect influential points is given by

$$\begin{aligned} DC_{it} &= \frac{1}{(p+q+r)}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(it)})^\top \mathbf{M}^\top \hat{\mathbf{W}} \mathbf{M}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(it)}) \\ &= \mathrm{r}^2_{SDit} \frac{\mathrm{c}_{it}\mathrm{h}_{it}}{(p+q+r)(1-\mathrm{h}_{it})^2}. \end{aligned} \tag{13}$$

where h_{it} is the t-th element of the main diagonal of $\mathbf{H}_i$, with $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Under homogeneity of the dispersion parameter, assume $r = 1$; otherwise, r is the dimension of the parametric vector $\boldsymbol{\delta}$.

Graphically, the plot of h_{it} versus $i-$ where h_{it} represents the t-th element of the main diagonal of $\mathbf{H}_i$, $i = 1, \ldots, n$ and $t = 1, \ldots, T-$ may be useful to identify leverage points. Widely discussed in the Statistics literature from [15], a relative high value of the Cook distance may indicate an influential point; to identify these points, do $(DC)_{it}$ versus the index i, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. At last, relative high values of $(\mathrm{r}_{PD})_{it}$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$, may indicate outliers.

4. Simulation study

A simulation study was conducted in order to evaluate the quality of the estimators obtained with the estimating equations proposed in this paper. The behavior of the estimators was examined under different sample sizes ($n = 50, 100, 500$), response vector sizes ($T = 2, 3, 5, 10$) and five different degrees of dependence.

Normal copulas (e.g. [21]) were used to generate multivariate data with $(0, 1)$ uniform marginal distributions. Let c_{it} be the generated copula value for the individual i in time t, the multivariate vector with ZAIG (or BEZI) marginal distribution was given by doing

$$y_{it} = \begin{cases} 0, & \text{if } c_{it} \leq \nu_{it}, \\ Q_{it}, & \text{if } c_{it} > \nu_{it}, \end{cases}$$

where y_{it} is the simulated value for individual i in time t and Q_{it} is the quantile of order c_{it} of the respective ZAIG (or BEZI) model.

The copulas were obtained from multivariate normal distributions with correlation coefficients ρ ($\rho = 0.0, 0.2, 0.5, 0.8$ and 0.9). In this context, ρ is a measure of the degree of dependence among the generated vector components. This procedure was done with the help of the packages COPULA [6] and GAMLSS [19] from R.

Each combination of n, T and ρ was simulated 10,000 times.

The model used in the simulation considered one independent variable, x_{it}, affecting ν_{it}, μ_{it} and σ_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$, where x_{it} were obtained by generating independent uniform distributions in the range $(-1, 1)$.

For ZAIG models

$$\nu_{it} = \frac{\exp(\beta_0 + \beta_1 \mathrm{x}_{it})}{1 + \exp(\beta_0 + \beta_1 \mathrm{x}_{it})}, \quad \mu_{it} = \exp(\gamma_0 + \gamma_1 \mathrm{x}_{it})$$

and

$$\sigma_{it} = \exp(\delta_0 + \delta_1 \mathrm{x}_{it}),$$

with $\beta_0 = \gamma_0 = \delta_0 = 0$, $\beta_1 = 2$ and $\gamma_1 = 1$ and $\delta_1 = 0.4$. The choice of $\delta_1 = 0.4$ minimized the amount of cases that failed to achieve convergence in the estimation process. The lack of convergence was sometimes due to extreme outliers in the data set. The value chosen for β_1 allows that the probability of zero response may vary between 12% and 88% roughly.

For BEZI models

$$\nu_{it} = \frac{\exp(\beta_0 + \beta_1 \mathrm{x}_{it})}{1 + \exp(\beta_0 + \beta_1 \mathrm{x}_{it})}, \quad \mu_{it} = \frac{\exp(\gamma_0 + \gamma_1 \mathrm{x}_{it})}{1 + \exp(\gamma_0 + \gamma_1 \mathrm{x}_{it})}$$

and

$$\sigma_{it} = \exp(\delta_0 + \delta_1 \mathrm{x}_{it}),$$

with $\beta_0 = \gamma_0 = \delta_0 = 0$, $\beta_1 = 2$ and $\gamma_1 = \delta_1 = 1$.

The relative mean absolute error (RMAE), obtained by dividing the observed mean absolute error by the parameter value, and the relative square root of the mean square error (RMSE), obtained by dividing the square root of the observed mean square error by the parameter value, were used to evaluate the simulation results. As the conclusion based on these two indicators were similar, only the RMAE results are presented in this section.

RMAE, in general, decrease, as n and T increase, for ZAIG and BEZI models. That is expected by the consistency of the estimators.

To make the conclusions of the study of the behavior of the errors due the dependency degree clear, for each combination of n and T, the ratio between RMAE for a fixed value of ρ and for $\rho = 0$ was calculated. Figures 1 and 2 illustrate these results. It may be seen that the effect of the correlation is low for small values of T, and it increases as T and ρ increase, for all n. This is more visible for values of ρ greater than 0.5. This conclusion is the same for ZAIG and BEZI models.

5. Application to the real data

In this section, traffic-related death rates in the southeastern cities of Brazil, between 2000 and 2002, will be analyzed. The dependent variables are the square-root of the annual mortality rate of traffic-related accidents per 100 thousand inhabitants (**Ratesq**) and the proportion of traffic-related deaths among all causes of death (**Proportion**).

We will build different models for each dependent variable; nevertheless, all models will use the following set of independent variables:

TABLE 1
The relative mean absolute error (RMAE) considering the **ZAIG** *simulation data set*

			ρ				
Parameter	n	T	0.0	0.2	0.5	0.8	0.9
β_1	50	2	18.9%	19.1%	19.0%	19.9%	20.4%
		3	15.1%	15.4%	15.5%	16.8%	17.3%
		5	11.6%	11.7%	12.3%	13.9%	14.5%
		10	8.0%	8.3%	9.1%	11.4%	12.3%
	100	2	13.0%	13.1%	13.4%	13.7%	13.9%
		3	10.5%	10.5%	11.0%	11.5%	12.0%
		5	8.0%	8.1%	8.7%	9.5%	10.2%
		10	5.7%	5.9%	6.5%	7.9%	8.7%
	500	2	5.7%	5.7%	5.8%	6.0%	6.1%
		3	4.6%	4.7%	4.7%	5.1%	5.3%
		5	3.6%	3.6%	3.8%	4.2%	4.5%
		10	2.5%	2.6%	2.9%	3.5%	3.8%
γ_1	50	2	22.9%	23.4%	23.1%	24.4%	25.0%
		3	18.5%	18.6%	19.0%	20.3%	21.0%
		5	14.1%	14.3%	15.1%	17.2%	18.3%
		10	9.6%	10.0%	11.5%	14.6%	15.9%
	100	2	15.6%	15.8%	15.8%	16.6%	16.9%
		3	12.7%	12.8%	13.2%	14.2%	14.8%
		5	9.8%	9.8%	10.6%	12.1%	12.9%
		10	6.8%	7.0%	8.0%	10.3%	11.4%
	500	2	6.8%	6.8%	7.1%	7.3%	7.5%
		3	5.5%	5.6%	5.9%	6.4%	6.5%
		5	4.3%	4.4%	4.7%	5.4%	5.8%
		10	3.1%	3.1%	3.7%	4.7%	5.2%
δ_1	50	2	44.6%	45.1%	45.5%	46.1%	46.8%
		3	35.4%	34.7%	35.3%	36.4%	37.4%
		5	26.3%	26.8%	27.2%	28.3%	29.8%
		10	18.3%	18.3%	18.9%	21.7%	23.3%
	100	2	29.4%	29.6%	29.7%	31.0%	30.6%
		3	23.5%	23.8%	24.3%	24.6%	25.8%
		5	17.9%	18.0%	18.4%	19.8%	20.7%
		10	12.7%	12.7%	13.4%	15.3%	16.4%
	500	2	12.6%	12.6%	12.9%	13.0%	13.3%
		3	10.3%	10.1%	10.4%	10.7%	11.1%
		5	7.9%	8.0%	8.2%	8.6%	9.1%
		10	5.6%	5.6%	5.9%	6.7%	7.2%

Year assumes the value 0 if the information is related to 2000, 1 if 2001 and 2 if 2002;

Lnpop natural logarithm of the number of the city's inhabitants, as determined by 2000 census;

Propurb proportion of the population living in the urban area of a municipality in 2000;

Table 2
The relative mean absolute error (RMAE) considering the **BEZI** *simulation data set*

			ρ				
Parameter	n	T	0.0	0.2	0.5	0.8	0.9
β_1	50	2	18.9%	19.1%	19.3%	19.7%	20.3%
		3	15.1%	15.2%	15.7%	16.6%	17.3%
		5	11.7%	11.8%	12.5%	13.8%	14.5%
		10	8.1%	8.2%	9.1%	11.1%	12.3%
	100	2	13.0%	13.1%	13.4%	13.8%	13.9%
		3	10.5%	10.6%	10.8%	11.7%	12.1%
		5	8.0%	8.3%	8.6%	9.7%	10.2%
		10	5.8%	5.8%	6.4%	7.9%	8.6%
	500	2	5.7%	5.7%	5.7%	6.0%	6.1%
		3	4.6%	4.7%	4.8%	5.1%	5.3%
		5	3.6%	3.6%	3.8%	4.3%	4.5%
		10	2.5%	2.6%	2.9%	3.4%	3.8%
γ_1	50	2	30.0%	30.4%	30.5%	30.4%	30.7%
		3	24.1%	24.0%	24.4%	24.4%	24.8%
		5	18.3%	18.4%	18.7%	19.0%	19.8%
		10	12.8%	12.9%	13.3%	14.5%	14.9%
	100	2	20.5%	20.7%	20.6%	20.5%	21.1%
		3	16.4%	16.6%	16.8%	16.8%	17.2%
		5	12.6%	12.8%	13.0%	13.3%	13.7%
		10	8.8%	9.1%	9.3%	10.0%	10.5%
	500	2	8.9%	8.9%	9.0%	9.2%	9.1%
		3	7.2%	7.3%	7.4%	7.5%	7.6%
		5	5.7%	5.7%	5.7%	6.0%	6.1%
		10	4.0%	4.0%	4.1%	4.5%	4.6%
δ_1	50	2	28.9%	29.7%	29.6%	30.1%	30.7%
		3	22.9%	22.7%	23.2%	23.8%	24.2%
		5	17.0%	17.2%	17.2%	18.5%	19.4%
		10	12.0%	11.9%	12.2%	13.6%	15.0%
	100	2	19.3%	19.4%	19.6%	19.9%	20.2%
		3	15.5%	15.5%	15.5%	15.7%	16.6%
		5	11.9%	11.9%	12.0%	12.7%	13.2%
		10	8.3%	8.3%	8.6%	9.7%	10.5%
	500	2	8.2%	8.4%	8.3%	8.4%	8.6%
		3	6.7%	6.7%	6.9%	7.0%	7.2%
		5	5.1%	5.1%	5.3%	5.6%	5.9%
		10	3.7%	3.7%	3.8%	4.3%	4.6%

Propmasc proportion of men in the city's population in 2000;
Prop2029 proportion of inhabitants aged 20–29 years; and
EHDI education index of the human development index of a municipality in 2000.

Table 3 shows some descriptive statistics for the response variables, there is a high incidence of zero values. By comparing the information for 2000, 2001 and

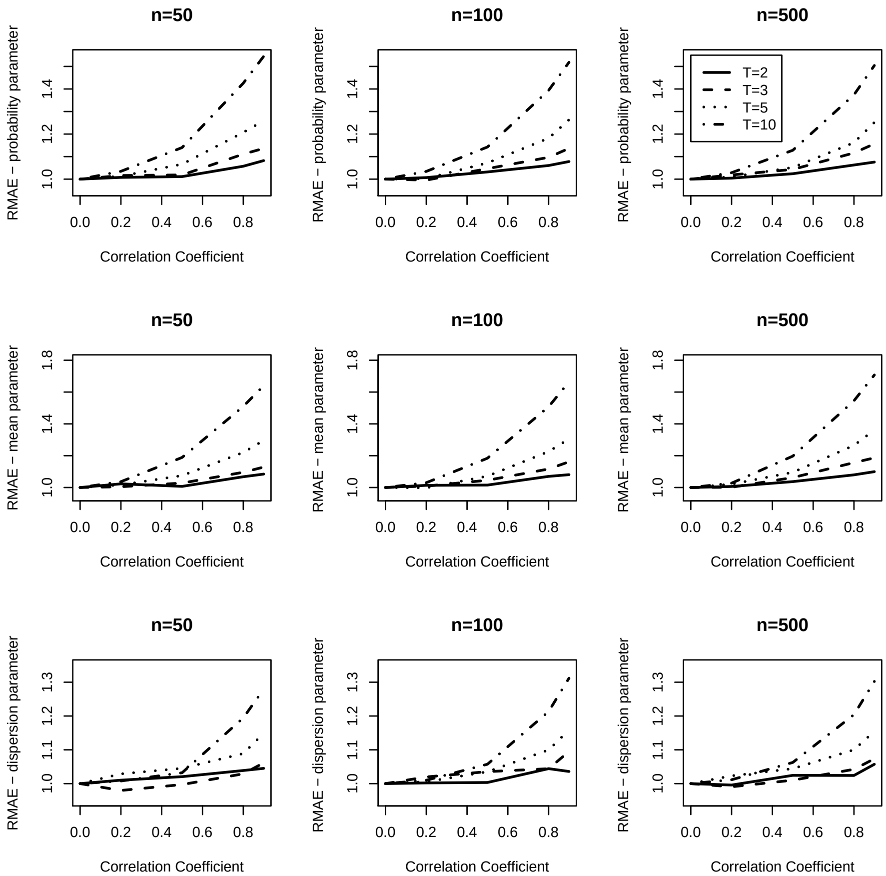

FIG 1. *Plot of the ratio between RMAE for ρ and RMAE when $\rho = 0$ considering the* **ZAIG** *simulation data set.*

2002, we notice, in both variables, a small variation of the proportion of municipalities with zero values, in the average values and in the standard deviation. This may be a suggestion time effect insistence. Due to some data problems, the sample sizes are slightly different.

The ZAIG distribution was used to model the square root of the traffic-related death rate and the BEZI distribution to model the percentage of traffic-related deaths. In both cases a model with heterogenous dispersion parameter was considered.

Let $\mathbf{x}_{it} = (1,\ \mathrm{Year01}_{it},\ \mathrm{Year02}_{it},\ \mathrm{Lnpop}_i,\ \mathrm{Propurb}_i,\ \mathrm{Propmasc}_i,\ \mathrm{Prop2029}_i,\ \mathrm{EHDI}_i)^\top$ is a covariate vector of the tth observation of the ith experimental unit, with $i = 1, \ldots, n$ and $t = 1$ (if Year= 2000), 2 (if Year= 2001) and 3 (if Year= 2002). Moreover, let $\mathrm{Year01}_{it}$ and $\mathrm{Year02}_{it}$ be indicator variables with value 1 when $t = 2$ and $t = 3$, respectively and $\mathbf{x}_{it} = \mathbf{d}_{it} = \mathbf{q}_{it}$.

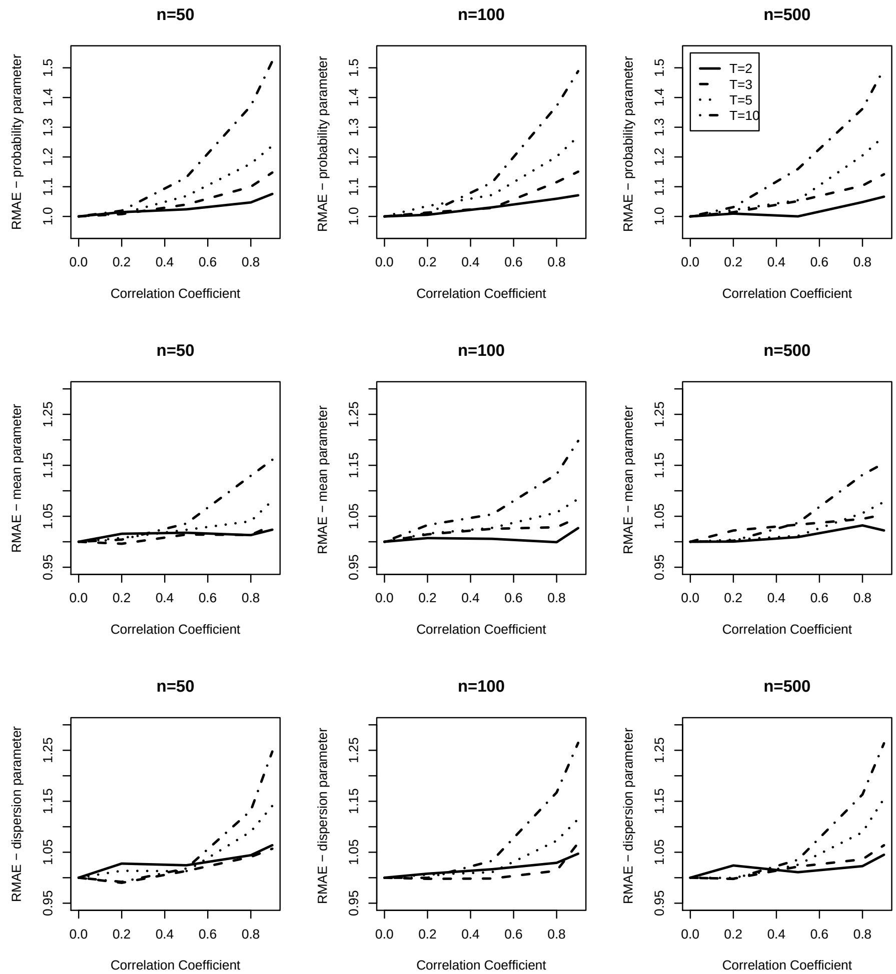

FIG 2. *Plot of the ratio between RMAE for ρ and RMAE when $\rho = 0$ considering the* **BEZI** *simulation data set.*

The estimates of $\boldsymbol{\theta}$ were obtained from the GAMLSS library available in *software* R; the standard error corrections and the diagnostic measures were obtained from a macro developed for R package.

5.1. Square-root of the annual traffic-related death rate

The following model was proposed in order to model the square root of the traffic-related death rate:

$$\nu_{it} = \frac{\exp(\mathbf{x}_{it}^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_{it}^\top \boldsymbol{\beta})}, \qquad \mu_{it} = \exp(\mathbf{x}_{it}^\top \boldsymbol{\gamma}) \qquad \text{and} \qquad \sigma_{it} = \exp(\mathbf{x}_{it}^\top \boldsymbol{\delta}).$$

TABLE 3

Descriptive statistics of the square-root of the annual traffic-related death rate and of the proportion of traffic-related deaths in the southeastern cities of Brazil

	Ratesq			Proportion		
Statistics	2000	2001	2002	2000	2001	2002
% of zeros	37.9	39.6	38.0	37.8	39.6	37.8
Minimum positive observation	1.0	1.0	1.2	0.2 %	0.1 %	0.2 %
Maximum	18.4	17.9	15.9	75 %	75%	80 %
Mean	3.0	2.9	3.0	4.6 %	4.3 %	4.7 %
Standard Deviation (SD)	2.8	2.8	2.8	7.6 %	7.4 %	7.9 %
Mean of the positive observations	4.8	4.8	4.9	7.3 %	7.2 %	7.6 %
SD of the positive observations	2.1	2.0	2.0	8.6 %	8.4 %	8.9 %
n		1665			1657	

TABLE 4

Parameter estimates described in $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)^\top$, of the standard error (SE) when there is supposition of independence among all observations (without correction) and when there is supposition of dependence among the observations of the same experimental unit (with correction), both modeled by **ZAIG** *distribution under heterogeneity of the dispersion parameter*

		WITHOUT correction			WITH correction		
$\boldsymbol{\beta}$	Estimate	SE	t	P-value	SE	t	P-value
Intercept	31.118	2.305	13.499	0.000	2.601	11.965	0.000
Year01	0.119	0.092	1.288	0.198	0.084	1.414	0.157
Year02	0.004	0.092	0.043	0.965	0.084	0.047	0.962
Lnpop	-1.588	0.061	-26.093	0.000	0.068	-23.53	0.000
Propurb	-0.253	0.261	-0.969	0.333	0.319	-0.793	0.428
Propmasc	-0.095	0.039	-2.409	0.016	0.046	-2.079	0.038
Prop2029	-0.093	0.031	-3.006	0.003	0.035	-2.634	0.009
IDHE	-13.163	0.981	-13.416	0.000	1.191	-11.056	0.000
$\boldsymbol{\gamma}$	Estimate	SE	t	P-value	SE	t	P-value
Intercept	0.843	0.463	1.820	0.069	0.618	1.363	0.173
Year01	-0.003	0.017	-0.186	0.852	0.012	-0.273	0.785
Year02	0.018	0.017	1.042	0.298	0.011	1.536	0.125
Lnpop	-0.130	0.009	-15.134	0.000	0.013	-10.302	0.000
Propurb	-0.312	0.058	-5.369	0.000	0.083	-3.741	0.000
Propmasc	0.004	0.008	0.471	0.637	0.012	0.330	0.741
Prop2029	0.005	0.006	0.808	0.419	0.010	0.543	0.587
IDHE	2.325	0.020	114.763	0.000	0.275	8.457	0.000
$\boldsymbol{\delta}$	Estimate	SE	t	P-value	SE	t	P-value
Intercept	-0.923	0.873	-1.057	0.291	1.050	-0.879	0.380
Year01	-0.009	0.031	-0.288	0.773	0.028	-0.320	0.749
Year02	-0.014	0.031	-0.435	0.663	0.028	-0.491	0.623
Lnpop	0.102	0.016	6.557	0.000	0.020	5.132	0.000
Propurb	-0.311	0.106	-2.95	0.003	0.123	-2.527	0.012
Propmasc	-0.032	0.016	-2.062	0.039	0.020	-1.571	0.116
Prop2029	0.013	0.012	1.081	0.280	0.016	0.822	0.411
IDHE	-0.215	0.381	-0.564	0.573	0.480	-0.447	0.655

The parameters estimates are presented in Table 4. The corrected results, considering the dependence between the observations, are presented in the last three columns of the table; in the three previous columns, the results that would be valid under total independence among the observations are presented. By

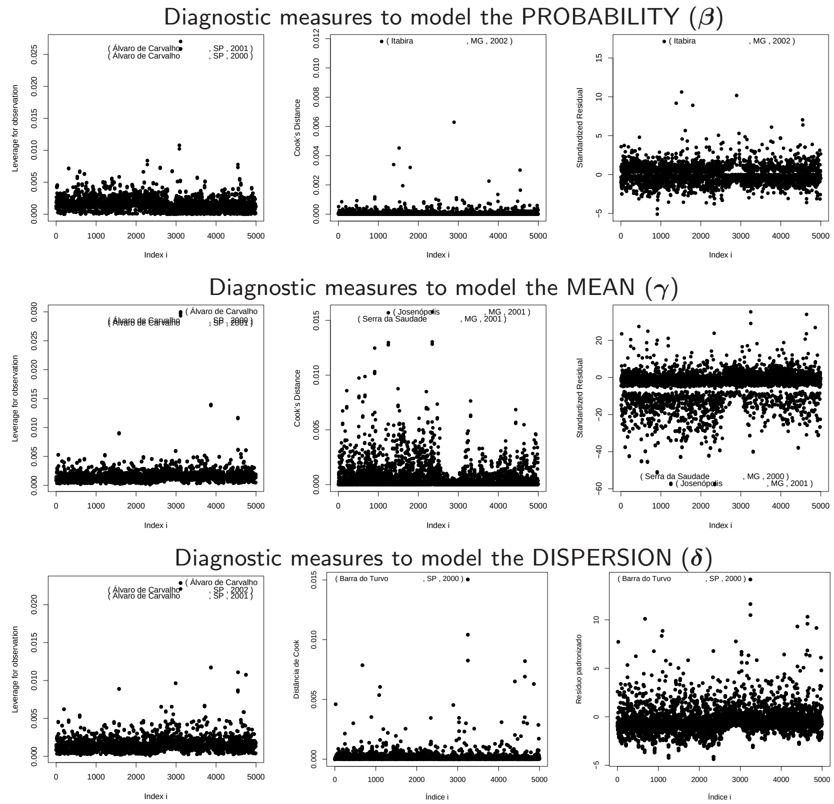

FIG 3. *Plot of the diagnostic measures (hat matrix, Cook's distance and standardized residual) obtained for each vector of the regression parameters: $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Regression model* **ZAIG** *under heterogeneity dispersion.*

comparing these columns, it is possible to notice the correction effect. The most important differences were observed in the intercept of μ_{it} ($\boldsymbol{\gamma}$) models and, for the coefficient of Propmasc, in the σ_{it} ($\boldsymbol{\delta}$) model.

In Figure 3, we may find the diagnostic measure graphs (projection matrix, Cook's distance and standardized residual) for each parameter vector of the regression models: $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. We may see that the city of Álvaro de Carvalho (SP) appears as a high leverage point for the three parameter vectors. The identification of the city of Álvaro de Carvalho as a high leverage point may be due the fact that it presents discrepant values for Propmasc (59.94, when the mean is 50.68 and the standard deviation is 1.25) and Prop2029 (24.95, when the mean is 16.54 and the standard deviation is 1.48).

For $\boldsymbol{\beta}$, the observation for the city of Itabira (MG), in 2002, appears as an influential point and as an outlier. Its observed value for Ratesq is zero and the model forecasts a low probability of assuming this value (0.45%); the estimated proportion of cities with actual zero value for this variable is 63.58%.

For $\boldsymbol{\delta}$, Barra do Turvo (SP), in 2000, appears as an influential point and as an outlier. This city presents high values for the response variables in the three years considered in the analysis (the highest in 2002, the second highest in 2000 and the third highest in 2001).

The observations for the city of Serra da Saudade (MG) and Josenópolis (MG), for $\boldsymbol{\gamma}$, in 2001, are highlighted in Cook's distance and standardized residual graphs. We did not find any explanation for this fact.

5.2. Proportion of traffic-related deaths

The regression models for the proportion of traffic-related deaths are

$$\nu_{it} = \frac{\exp(\mathbf{x}_{it}^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_{it}^\top \boldsymbol{\beta})}, \qquad \mu_{it} = \frac{\exp(\mathbf{x}_{it}^\top \boldsymbol{\gamma})}{1+\exp(\mathbf{x}_{it}^\top \boldsymbol{\gamma})} \qquad \text{and} \qquad \sigma_{it} = \exp(\mathbf{x}_{it}^\top \boldsymbol{\delta}).$$

The parameter estimates may be found in Table 5. As we have seen previously, the correction effect occurs both increasing and decreasing the standard errors.

TABLE 5

Parameter estimates described in $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)^\top$, of the standard error (SE) when there is supposition of independence among all observations (without correction) and when there is supposition of dependence among the observations of the same experimental unit (with correction), both modeled by **BEZI** *distribution under heterogeneity of dispersion parameter*

		WITHOUT correction			WITH correction		
$\boldsymbol{\beta}$	Estimate	SE	t	P-value	SE	t	P-value
Intercept	31,498	2.319	13.583	0.000	2.621	12.018	0.000
Year01	0.133	0.093	1.430	0.153	0.084	1.583	0.114
Year02	0.000	0.093	0.000	1.000	0.085	0.000	1.000
Lnpop	-1.601	0.061	-26.246	0.000	0.068	-23.544	0.000
Propurb	-0.278	0.262	-1.061	0.289	0.321	-0.866	0.387
Propmasc	-0.099	0.040	-2.475	0.013	0.046	-2.152	0.032
Prop2029	-0.098	0.031	-3.161	0.002	0.036	-2.722	0.007
IDHE	-13.152	0.986	-13.339	0.000	1.198	-10.978	0.000
$\boldsymbol{\gamma}$	Estimate	SE	t	P-value	SE	t	P-value
Intercepto	-5.792	0.985	-5.880	0.000	1.625	-3.564	0.000
Year01	0.000	0.034	-0.009	0.993	0.029	-0.010	0.992
Year02	0.051	0.034	1.500	0.134	0.028	1.821	0.069
Lnpop	-0.428	0.017	-25.176	0.000	0.042	-10.190	0.000
Propurb	-0.693	0.122	-5.680	0.000	0.210	-3.300	0.001
Propmasc	0.067	0.018	3.722	0.000	0.030	2.233	0.026
Prop2029	0.107	0.014	7.643	0.000	0.023	4.652	0.000
IDHE	3.176	0.435	7.301	0.000	0.725	4.381	0.000
$\boldsymbol{\delta}$	Estimate	SE	t	P-value	SE	t	P-value
Intercept	1.351	1.792	0.754	0.451	3.040	0.444	0.657
Year01	0.027	0.065	0.415	0.678	0.070	0.386	0.700
Year02	-0.014	0.064	-0.219	0.827	0.071	-0.197	0.844
Lnpop	0.536	0.032	16.750	0.000	0.089	6.022	0.000
Propurb	1.081	0.217	4.982	0.000	0.379	2.852	0.004
Propmasc	-0.014	0.032	-0.438	0.662	0.057	-0.246	0.806
Prop2029	-0.070	0.025	-2.800	0.005	0.045	-1.556	0.120
IDHE	-2.779	0.791	-3.513	0.000	1.334	-2.083	0.037

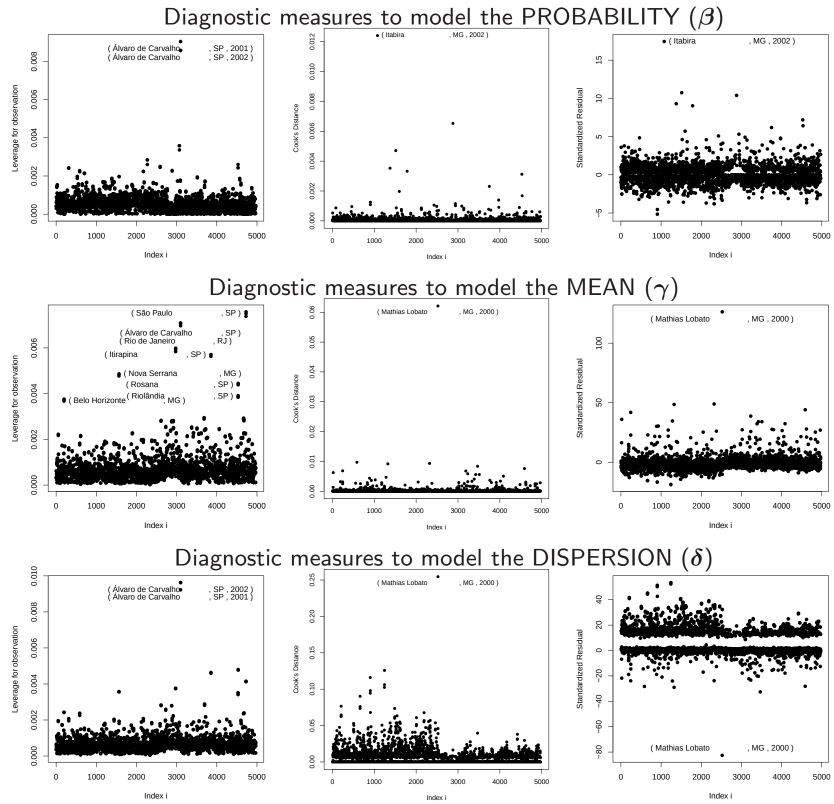

FIG 4. *Plot of the diagnostic measures (hat matrix, Cook's distance and standardized residual) obtained for each vector of the regression parameters: $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Regression model* **BEZI** *under heterogeneity dispersion.*

Figure 4 brings the diagnostic measure graphs for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$.

Three state capitals (out of four) – São Paulo, Rio de Janeiro and Belo Horizonte – appear as leverage points in the mean modeling. This may be explained by high value of the variable Lnpop. In 2000, the mean population value of the cities included in the analysis were 12.2 thousand inhabitants, while the population of these capitals were 10.5 million, 5.9 million and 2.2 million, respectively.

Álvaro de Carvalho was identified as a leverage point, just like in the last section. Itirapina, Riolândia and Rosana were identified as high leverage points too; the first two presented a high value for the variable Propmasc (Itaparina: 55.67 and Riolândia: 55.76). Rosana showed low value for Propurb, 0.26) (the mean was 0.70 and the standard deviation was 0.21), and high value for EHDI 0.91 (the mean was 0.82 and the standard deviation was 0.06); this behavior is unexpected since the correlation between these variables is 0.72. Nova Serrana was identified as a high leverage point, but we did not find any explanation for this fact.

The city of Itabira (MG), in 2002, also appears as an influential point and as an outlier to BEZI model. The value for Proportion in 2002 is zero, but the model also forecasts a low probability of assuming this value, 0.42%.

Mathias Lobato is detected as outlier and as influent for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. It is a city with a low population (Lnpop=8.0, while the mean population is 9.4 with standard deviation 1.2) and, in 2000, the observed proportion of traffic-related deaths was extremely high: 50%. Besides its value for Propurb was 90%, what is unexpected for a city with 3,642 inhabitants. The correlation between Lnpop and Propurb is 0.49.

6. Concluding remarks

In this paper we proposed estimating equations for regression models for zero-inflated random variables. In particular, we focused on BEZI and ZAIG distributions. The estimating equations are similar to the score function obtained in the full independence case. In practical terms, the method provides a correction for the standard errors of the parameter estimators.

The results are directly applicable to the analysis of any zero-inflated semi-continuous response variable. In particular, to zero-inflated Gamma (ZIG), zero-inflated log-normal (ZILN) and zero-inflated truncated Pareto (ZITPo), see [12] and [1] for further details about these distributions.

Furthermore we proposed diagnostic techniques to identify outliers, leverage points and Cook's influence points, which is an advance compared to other studies cited in this paper.

The simulation study suggests that the estimating errors decrease as the sample size and the size of the response vector grow and increase when the dependence degree among the response vector components is high, mainly for large samples and response vector dimensions.

We applied the methods to two data sets to get a correct value for the standard error estimates of regression model parameters, in the presence of dependency among observations of the same sample unit. The standard errors obtained from Fisher's information acquired under the assumption of independence among all observations would lead to wrong inferences.

References

[1] Couturier, D.-L. and Victoria-Feser, M.-P. (2010). Zero-inflated truncated generalized Pareto distribution for the analysis of radio audience data. *Annals of Applied Statistics* **4**(4) 1824–1846. MR2829937

[2] Dobbie, M. J. and Welsh, A. H. (2001). Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics* **43** 431–444. MR1872202

[3] Gan, N. (2000). *General zero-inflated models and their applications*. Thesis (Ph.D.), North Carolina State University. 132 pp. MR2702750

[4] Hall, D. B. and Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling* **4** 161–180. MR2062098

[5] Heller, G., Stasinopoulos, M. and Rigby, B. (2006). *The zero-adjusted inverse gaussian distribution as a model for insurance claims*. Proceedings of the 21th International Workshop on Statistical Modelling, Ireland-Galway. http://studweb.north.londonmet.ac.uk/~stasinom/papers/ZAIG.pdf.

[6] Hofert, M., Kojadinovic, I., Maechler, M. and Yan, J. (2014). *Copula: Multivariate Dependence with Copulas*. R package version 0.999-10. http://CRAN.R-project.org/package=copula.

[7] Jong, P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press.

[8] Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London. MR1462891

[9] Jørgensen, B. and Labouriau, R. S. (1994). *Exponential Families and Theoretical Inference*. Lecture Notes, Department of Statistical, University of British Columbia.

[10] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

[11] Martinez, R. O. (2008). *Modelos de regressão beta inflacionados*. Thesis (Ph.D.). IME-USP, São Paulo.

[12] Mills, E. D. (2013). *Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data*. Thesis (Ph.D.), Graduate College of The University of Iowa. 280 pp. http://ir.uiowa.edu/etd/2583/. Acessed in 2014/07/29.

[13] Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5** 1–19. MR2133525

[14] Ospina, R. and Ferrari, S. L. P. (2010). Inflated beta distribution. *Stat Papers* **51** 111–126. MR2556590

[15] Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics* **9** 705–724. MR0619277

[16] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[17] Ridout, M. S., Demétrio, C. G. B. and Hinde, J. P. (1998). *Models for count data with many zeros*. Proceedings of the XIXth International Biometrics Conference. Cape Town.

[18] Song, X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer, New York. MR2377853

[19] Stasinopoulos, M., Rigby, B. and Akantziliotou, C. (2008). *Instructions on How to Use the Gamlss Package in R*. 2nd edition. http://studweb.north.londonmet.ac.uk/~stasinom/papers/gamlss-manual.pdf.

[20] Venezuela, M. K., Botter, D. A. and Sandoval, M. C. (2007). Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation* **77** 879–888. MR2409950

[21] Yan, J. (2007). Enjoy the Joy of Copulas: With a Package Copula. *Journal of Statistical Software* **21**(4) 1–21. http://www.jstatsoft.org/v21/i04/