Brazilian Journal of Biometrics, **43**, 1–21 (2025)
doi:10.28951/bjb.v43i1.728

BRAZILIAN JOURNAL OF BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Residual analysis for discrete correlated data in the multivariate approach

Lizandra C. Fabio,[1] Cristian Villegas,[2] Abu Sayed Md. Al Mamun,[3] and Jalmar M F Carrasco⋆,[1]

[1] Department of Statistics, Federal University of Bahia, Salvador - BA, Brazil
[2] University of São Paulo, São Paulo - SP, Brazil
[3] Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh
⋆Corresponding author. Email: carrascojalmar@gmail.com

(Received: March 4, 2024; Revised: June 11, 2024; Accepted: June 14, 2024; Published: February 11, 2025)

Abstract

The residual distributions obtained from discrete correlated and uncorrelated data cannot be well approximated to the standardized normal distribution. In this case, the efficiency in checking the adequacy of the model to the data and detecting outliers is not guaranteed. Thus, alternative measures for residual analysis have been considered in several classes of models and their properties have been assessed. In this paper, we investigate the empirical distribution of four residuals of the multivariate negative binomial regression (MNBR) model. In our study, we propose standardized weighted and standardized Pearson residuals; we also consider the standardized component of deviance and quantile residuals suggested by Fabio *et al.* (2012) and Fabio *et al.* (2023), respectively. Monte Carlo simulation results reveal that the concordance of the empirical distribution of the residuals to the standard normal distribution depends on the dispersion parameter. Furthermore, the impact on residual analysis when the random effect distribution is misspecified is explored. We concluded that the quantile and standardized weighted residuals presented better performances.

Keywords: Count data; Monte Carlo simulation; Multivariate Negative Binomial distribution; Overdispersion; Residual analysis.

1. Introduction

Residual analysis is used to assess the adequacy of models and detect outliers (Hardin & Hilbe, 2016). It is desirable that the residual distribution follows the standardized normal distribution to check the lack of fitting through graphical methods. However, this assumption cannot be guaranteed when the response variable is nonnormal. Based on this fact, some Monte Carlo simulation studies have been carried out for both correlated and uncorrelated data to assess the properties of the standardized deviance component, Pearson residuals and alternative residuals when the random variable is discrete or positive continuous. Espinheira *et al.* (2008) presented the weighted and standardized weighted residuals to detect the misspecification of a class of beta regression models. Numerical procedures showed that the last residual is better approximated to the standard normal distribution. The authors also verified that the standardized weighted residual is able to clearly identify atypical and influential observations. Scudilio & Pereira (2020) demonstrated using Monte Carlo simulation studies that the adjusted quantile residual computed from gamma and inverse normal distribution outperformed other residuals. Feng *et al.* (2020) revealed from their simulation results that the distributions of the Pearson and standardized deviance component residuals obtained from Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial are right-skewed and heavy-tailed in comparison to the normal distribution. Pereira *et al.* (2020) introduced a class of residuals for checking the overall adequacy of a zero-adjusted regression model, which is superior for detecting outliers compared to the randomized quantile residual. Fabio *et al.* (2012) and Fabio *et al.* (2023) have suggested an extension of these kind of residuals for the residual analysis of the MNBR model. The MNB distribution was deduced by Fabio *et al.* (2012) from the random intercept Poisson mixed model where the random effect is assumed to follow the GLG distribution (Lawless, 1987). The GLG distribution can be skewed to the right or to the left, and the normal distribution is a particular case. Further, the MNB distribution belongs to the discrete multivariate exponential family (Johnson *et al.*, 1997) . Its dispersion parameter depends on the shape parameter of the GLG distribution. Fabio *et al.* (2023) showed that this parameter has the desirable asymptotic consistency when the shape parameter assumes small values. This parameter can have an effect on the residual analysis of the MNBR model. We aim to investigate the empirical distribution of four residuals of the MNBR model. We propose the standardized weighted and standardized Pearson residuals, and consider standardized component of deviance and quantile residuals suggested by Fabio *et al.* (2012) and Fabio *et al.* (2023), respectively. Under the assumption of MNBR model misspecification, Monte Carlo simulations are performed to evaluate the approximation of the empirical distribution of residuals with respect to the standard normal distribution and their behavior.

This paper is organized as follows: In Section 2, we define the MNBR model and present its properties. In Section 3, we propose standardized weighted and standardized Pearson residuals. Additionally, we consider the standardized component of deviance and quantile residuals. In Section 4, we perform a simulation study to evaluate the behavior of the empirical distributions of the residuals under different scenarios. In Section 5, the residual analysis is applied to two real data sets. Finally, in Section 6, we discuss the conclusions.

2. MNBR model

In this section, we present the MNB distribution which was deduced from a random intercept Poisson model. The generalized log-gamma is assumed by Fabio *et al.* (2012) as the distribution for random effect. Let b be a random variable following a generalized log-gamma (GLG) distribution, for which probability density function (pdf) is given by

$$f(b;\mu,\sigma,\lambda)=\begin{cases}\frac{c(\lambda)}{\sigma}\exp\left[\frac{(b-\mu)}{\lambda\sigma}-\frac{1}{\lambda^2}\exp\left\{\frac{\lambda(b-\mu)}{\sigma}\right\}\right], & \text{if} \quad \lambda \neq 0,\\ \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(b-\mu)^2}{2\sigma^2}\right\}, & \text{if} \quad \lambda = 0,\end{cases} \tag{1}$$

with $b \in \mathbb{R}$. The parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\lambda \in \mathbb{R}$ are the location, scale, and shape parameters, respectively, and $c(\lambda) = |\lambda|(\lambda^{-2})^{\lambda^{-2}}/\Gamma(\lambda^{-2})$ with $\Gamma(\cdot)$ being the gamma function. We denoted $b \sim \text{GLG}(\mu,\sigma,\lambda)$. The extreme value distribution is a particular case of (1) when $\lambda = 1$. For $\lambda < 0$ the GLG pdf of b is skewed to the right and for $\lambda > 0$ it is skewed to the left. For $\lambda = 0$, (1) reduces to normal distribution. Figure 1 shows the GLG distribution for $\lambda = 2$ (skewed to the right), $\lambda = -2$ (skewed to the left) and $\lambda = 0$ (symmetrical).

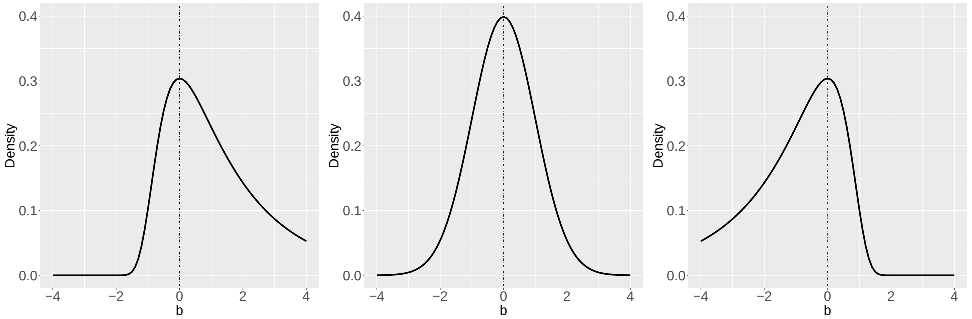

Figure 1. The probability density function of the generalized log-gamma distribution for the parameter λ=-2 (left), λ=0 (center) and λ=2 (right).

Fabio *et al.* (2012) showed that the MNB distribution can be derived from the random intercept Poisson-GLG model by assuming that $\sigma = \lambda$ and $\mu = 0$, in expression (1). Furthermore, the authors consider the reparametrization, $\phi = \lambda^{-2}$. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$ where each y_{ij} denotes the jth measurement taken on the ith subject or cluster, for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. The marginal probability mass function (pmf) of $\boldsymbol{y}_i$ is given by

$$f(\boldsymbol{y}_i; \theta) = \frac{\Gamma(\phi + y_{i+})\phi^{\phi}}{\left(\prod_{j=1}^{m_i} y_{ij}!\right)\Gamma(\phi)} \frac{\exp\left(\sum_{j=1}^{m_i} y_{ij}\log(\mu_{ij})\right)}{(\phi + \mu_{i+})^{\phi + y_{i+}}}, \tag{2}$$

where $\theta = (\beta^\top, \phi)^\top$, $\phi = \lambda^{-2}$ is the dispersion parameter, ϕ^{-1} is the overdispersion parameter, $\Gamma(\cdot)$ is the gamma function, $y_{i+} = \sum_{j=1}^{m_i} y_{ij}$ and $\mu_{i+} = \sum_{j=1}^{m_i} \mu_{ij}$. Its marginals are negative binomial (NB) distributions, with means $\mathrm{E}(y_{ij}) = \mu_{ij}$, variances $\mathrm{Var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2/\phi$, covariances $\mathrm{Cov}(y_{ij}, y_{ij'}) = \mu_{ij}\mu_{ij'}/\phi$ for $j \neq j'$, and intraclass correlation $\mathrm{Corr}(y_{ij}, y_{ij'}) = \sqrt{\mu_{ij}\mu_{ij'}}/\sqrt{(\phi + \mu_{ij})(\phi + \mu_{ij'})}$ for $j \neq j'$ is always positive. We can easily see that when ϕ assumes large values, the marginals of MNB distribution behave approximately as independent Poisson distribution with mean μ_{ij}. Otherwise, when ϕ assumes values close to zero, the intraclass correlation tends to one. We denote by $\boldsymbol{y}_i \sim \mathrm{MNB}(\boldsymbol{\mu}_i, \phi)$, independent vectors of random outcomes which follow the probability function given in (2) with $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^\top$, and $\phi > 0$. The MNB regression (MNBR) model follows the hierarchical structure, that is, (i) $\boldsymbol{y}_i \overset{\text{ind}}{\sim} \mathrm{MNB}(\boldsymbol{\mu}_i, \phi)$ and (ii) $\log(\mu_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$, with $\boldsymbol{x}_{ij} = (\boldsymbol{x}_{ij1}, \boldsymbol{x}_{ij2}, \ldots, \boldsymbol{x}_{ijp})^\top$ containing values of the explanatory variables for the ith cluster (subject), and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is the vector of regression coefficients. The inferences for the MNBR model are obtained from expression (3), in which its formulation has a closed form. Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top$ be the vector containing all measured outcomes for the ith subject, the log-likelihood function is given by

$$\begin{aligned} \ell(\theta) = \sum_{i=1}^{n} \log\left\{\frac{\Gamma(\phi + y_{i+})}{\Gamma(\phi)}\right\} - \sum_{i=1}^{n}\sum_{j=1}^{m_i} \log\left(y_{ij}!\right) + n\phi\log(\phi) - \\ \phi\sum_{i=1}^{n}\log(\phi + \mu_{i+}) + \sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}\log\left\{\frac{\mu_{ij}}{\phi + \mu_{i+}}\right\}, \end{aligned} \tag{3}$$

where $\theta = (\beta^\top, \phi)^\top$. The maximum likelihood (ML) estimates $\widehat{\theta}$ of θ are obtained following the Fisher's scoring iterative algorithm described by Fabio *et al.* (2012). The estimate of the β and ϕ parameter could be obtained by

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} + (X^\top W^{(t)} X)^{-1} X^\top (y - \mu^{*(t)}),\\ \phi^{(t+1)} &= \phi^{(t)} - \frac{U_\phi(\phi^{(t)})}{L_{\phi\phi}(\phi^{(t)})},\end{aligned}$$

with

$$U_\phi = \sum_{i=1}^{n} \left\{ \sum_{j=0}^{y_{i+}-1} (j+\phi)^{-1} - \frac{y_{i+}}{\phi + \mu_{i+}} - \log(1 + \phi^{-1}\mu_{i+}) + \frac{\mu_{i+}}{\phi + \mu_{i+}} \right\},$$

where the inner summation is 0 when $y_{i+} - 1 < 0$ and

$$L_{\phi\phi} = \frac{\partial U_\phi}{\partial \phi} = \sum_{i=1}^{n} \left\{ \sum_{j=0}^{(y_{i+}-1)} \frac{y_{i+}}{(\phi + \mu_{i+})^2} + \frac{\phi^{-1}\mu_{i+}}{\phi + \mu_{i+}} - \frac{\mu_{i+}}{(\phi + \mu_{i+})^2} - (j+\phi)^{-2} \right\},$$

for the $t = 0, 1, 2, \ldots$, $X^\top = (X_1, \ldots, X_n)^\top$ with $X_i = (x_{i1}, \ldots, x_{im_i})^\top$ and $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$; $y = (y_1^\top, \ldots, y_n^\top)^\top$ with $y_i = (y_{i1}, \ldots, y_{im_i})^\top$; $\mu^* = (\mu_1^*, \mu_2^*, \ldots, \mu_n^*)^\top$ with $\mu_i^* = a_i \odot \mu_i$, $a_i = (\phi + y_{i+})/(\phi + \mu_{i+})$, $\mu_i = (\mu_{i1}, \ldots, \mu_{im_i})^\top$ and $\odot$ define the Hadamard product; $W = \text{diag}(W_1, \ldots, W_n)$ with $W_i = \text{diag}(\mu_i) - (\phi + \mu_{i+})^{-1}\mu_i\mu_i^\top$ and diag$(\cdot)$ represent the diagonal matrix. At convergence, we obtain

$$\widehat{\beta} = (X^\top \widehat{W} X)^{-1} X^\top (y - \widehat{\mu}^*),$$

where the elements from $\widehat{W}$ are $\widehat{W}_i = \text{diag}(\widehat{\mu}_i) - (\widehat{\phi} + \widehat{\mu}_{i+})^{-1}\widehat{\mu}_i\widehat{\mu}_i^\top$, the elements from $\widehat{\mu}^*$ are $\widehat{\mu}_i^* = \widehat{a}_i \odot \widehat{\mu}_i$, $\widehat{a}_i = (\widehat{\phi} + y_{i+})/(\widehat{\phi} + \widehat{\mu}_{i+})$ and $\widehat{\phi}$ is the maximum likelihood estimate. Under standard regularity conditions, assuming $\widehat{\phi}$ is a consistent estimator, the approximate distribution of $(\widehat{\beta} - \beta)$ is the multivariate normal distribution with mean zero and covariance matrix $I_{\beta\beta} = -\text{E}[\partial \ell^2(\theta)/\partial\beta\partial\beta^\top] = X^\top W X$.

3. Residual analysis

In this section, we propose the standardized Pearson and standardized weight as measures of agreement between the data and the fitted MNBR model. We also review other kinds of residuals suggested by Fabio *et al.* (2012) and Fabio *et al.* (2023). We deduce the standardized weighted residual from the Fisher scoring iterative process of β when ϕ is fixed, similarly to GLMs theory (Agresti, 2015; Faraway, 2016). Since the convergence of this estimation method is satisfied, the $\widehat{\beta}$ estimator from the MNBR model reduces to

$$\widehat{\beta} = (X^\top \widehat{W} X)^{-1} X^\top \widehat{W} z, \tag{4}$$

where $z = X\widehat{\beta} + \widehat{W}^{-1}(y - \widehat{\mu}^*)$. Despite z differing from the usual form presented in GLMs, the estimator of β arises with equivalent formulation. Hence, the weighted residual vector based on (4) is expressed by $r = \widehat{W}^{1/2}(z - X\widehat{\beta}) = \widehat{W}^{-1/2}(y - \widehat{\mu}^*)$. The weighted residual for the ith subject and

ijth observation are given by $\boldsymbol{r}_i = \widehat{\boldsymbol{W}}_i^{-1/2}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i^*)$ and $r_{ij}^w = (y_{ij} - \widehat{\mu}_{ij}^*)/\sqrt{\widehat{w}_i^{jj}}$, respectively. The w_i^{jj} represent the jth measurement taken on the ith subject of the $\boldsymbol{W}$ matrix . Furthermore, as

$$\mathrm{Cov}(\boldsymbol{r}) = [\widehat{\boldsymbol{W}}^{1/2} - \widehat{\boldsymbol{W}}^{1/2}X(X^\top \widehat{\boldsymbol{W}}X)^{-1}X^\top \widehat{\boldsymbol{W}}] \times \mathrm{Cov}(\boldsymbol{z}) \times$$
$$[\widehat{\boldsymbol{W}}^{1/2} - \widehat{\boldsymbol{W}}^{1/2}X(X^\top \widehat{\boldsymbol{W}}X)^{-1}X^\top \widehat{\boldsymbol{W}}]^\top,$$

it is possible to conclude that $\mathrm{Cov}(\boldsymbol{z}) \approx \widehat{\boldsymbol{W}}^{-1}$ and $\mathrm{Cov}(\boldsymbol{r}) = (\boldsymbol{I}_N - \widehat{\boldsymbol{H}})$ where $\boldsymbol{I}_N$ is an identity matrix of order N and the weighted hat matrix is given as

$$\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{1/2}X(X^\top \widehat{\boldsymbol{W}}X)^{-1}X^\top \widehat{\boldsymbol{W}}^{1/2}.$$

The standardized weighted residual vector which is orthogonal to the weighted fitted linear predictor, $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{W}}^{1/2}X\widehat{\boldsymbol{\beta}}$, is defined by $\boldsymbol{r}^p = (\boldsymbol{I} - \widehat{\boldsymbol{H}})^{-1/2}\widehat{\boldsymbol{W}}^{-1/2}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}^*)$. The standardized weighted residual corresponding to ith subject takes the form $\boldsymbol{r}_i^p = (\boldsymbol{I}_n - \widehat{\boldsymbol{H}}_i)^{-1/2}\widehat{\boldsymbol{W}}_i^{-1/2}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i^*)$, where $\widehat{\boldsymbol{H}}_i = \widehat{\boldsymbol{W}}_i^{1/2}X_i(X_i^\top \widehat{\boldsymbol{W}}_i X_i)^{-1}X_i^\top \widehat{\boldsymbol{W}}_i^{1/2}$. In addition, the residual to ijth observation is given as

$$r_{ij}^{sw} = \frac{y_{ij} - \widehat{\mu}_{ij}^*}{\sqrt{\widehat{w}_i^{jj}(1 - \widehat{h}_i^{jj})}}, \tag{5}$$

where h_i^{jj} represents the jth measurement taken on the ith subject of the $\boldsymbol{H}$ matrix, for $i = 1, 2, \ldots, n$ and $j = 1, \ldots, m_i$.

Waller & Zelterman (1997) suggested the Pearson residuals, for the ijth observation to assess the adequacy of the MNBR model, which is defined as $r_{ij}^p = (y_{ij} - \widehat{\mu}_{ij})/\sqrt{\widehat{\mu}_{ij}(1 + \widehat{\phi}^{-1}\widehat{\mu}_{ij})}$ by assuming that ϕ is fixed. We propose the standardized Pearson residuals considering the leverages of the observations for the standardization of r_{ij}^p. Thus, its expression takes the following form,

$$r_{ij}^{sp} = \frac{r_{ij}^p}{\sqrt{1 - \widehat{h}_i^{jj}}} = \frac{(y_{ij} - \widehat{\mu}_{ij})}{\sqrt{\widehat{\mu}_{ij}(1 + \widehat{\phi}^{-1}\widehat{\mu}_{ij})}\sqrt{1 - \widehat{h}_i^{jj}}}. \tag{6}$$

Fabio *et al.* (2012) defined the standardized deviance component as residuals for the ith subject (or cluster) by considering that ϕ is fixed . Its goodness-of-fit measure can be expressed as $r_i^d = d(\boldsymbol{y}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\phi})/\sqrt{1 - \widehat{h}_i}$, where $d(\boldsymbol{y}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\phi}) = \pm\sqrt{2}\{d^2(\boldsymbol{y}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\phi})\}^{1/2}$ with

$$d^2(\boldsymbol{y}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\phi}) = \widehat{\phi}\log\left(\frac{\widehat{\phi} + \widehat{\mu}_{i+}}{\widehat{\phi} + y_{i+}}\right) + \sum_{\{\forall y_{ij} \neq 0\}} y_{ij}\log\left(\frac{y_{ij}(\widehat{\phi} + \widehat{\mu}_{i+})}{\widehat{\mu}_{ij}(\widehat{\phi} + y_{i+})}\right), \tag{7}$$

where $h_i = \sum_{j=1}^{m_i} h_i^{jj}$ is a component of $\boldsymbol{H}_i$ matrix. The sign being the same as that of $(y_{i+} - \widehat{\mu}_{i+})$. Recently, Fabio *et al.* (2023), under the supposition that y_{i+} follows a negative binomial distribution (Tsui, 1986), employed the randomized quantile residual, r_i^q, for detecting outliers of the ith subject in the MNBR model. If $F(y_{i+}; \mu, \phi)$ is the cumulative distribution of $f(y_{i+})$, $a_i = \lim_{y \uparrow y_{i+}} F(y; \widehat{\mu}_i, \widehat{\phi})$, and $b_i = F(y_{i+}; \widehat{\mu}_i, \widehat{\phi})$, then the randomized quantile residuals for $\boldsymbol{y}_i$ are given by $r_{q,i} = \Phi^{-1}(u_i)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal and u_i is a uniform random variable on the interval $(a_i, b_i]$.

In general, it is recommended to draw the normal probability plot with an envelope for detecting possible departures from the assumptions as well as outlying observations or subjects. However, simulation studies are conducted with the aim of identifying their best performance.

4. Numerical results

In this section, Monte Carlo simulation studies are conducted to evaluate the approximation of the empirical distributions of the standardized weighted, standardized Pearson, quantile, and standardized deviance component residuals with respect to the standard normal distribution. The numerical results are based on R = 10,000 Monte Carlo replications. We consider two scenarios, in which the response variable $\boldsymbol{y} = (\boldsymbol{y}_i^\top, \dots, \boldsymbol{y}_n^\top)$, where $\boldsymbol{y}_i \sim \text{MNB}(\boldsymbol{\mu}_i, \phi)$ is generated following a hierarchical structure, $(i) y_{ij}|b_i \overset{\text{ind}}{\sim} \text{Poisson}(u_{ij})$, $(ii) \log(u_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_i$, and $(iii) b_i \overset{\text{iid}}{\sim} \text{GLG}(0, \lambda, \lambda)$. The covariate values are obtained as random draws in the following distributions: $x_{ij1} \sim N(0, 1)$ and $x_{ij2} \sim \mathit{Bernoulli}(0.5)$ for $i = 1, \dots, n$ and $j = 1, 2, 3$. The covariate values remain constant throughout the simulation. Further, it is assumed that $\boldsymbol{\beta} = (1.5, 1.0, 0.0)^\top$ and $\lambda = \phi^2$ in the GLG distribution with $\phi(\lambda) = 0.5(1.41)$, $1.0(1.0)$, $3.0(0.58)$, $20(0.22)$, $100(0.1)$ and $500(0.04)$. A three-simulation study is performed to evaluate the impact of misspecifying the random effect distribution on the residuals. For samples size $n = 100$ and 200, the vector $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_n^\top)^\top$ is generated wrongly from a random intercept Poisson-Normal distribution following a hierarchical structure $(i) y_{ij}|b_i \overset{\text{ind}}{\sim} \text{Poisson}(u_{ij})$, $(ii) \log(u_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_i$, and $(iii) b_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma)$, with $\sigma = \phi^{-1} = 2.0, 0.33, 0.01$ and fitted under MNBR model.

4.1 Scenario 1

In this scenario, the empirical distribution of (5), (6), (7) and quantile residuals is evaluated and compared with the quantiles of the standard normal distribution. We consider six different values for the dispersion parameter and specify a sample size as $n = 30$.

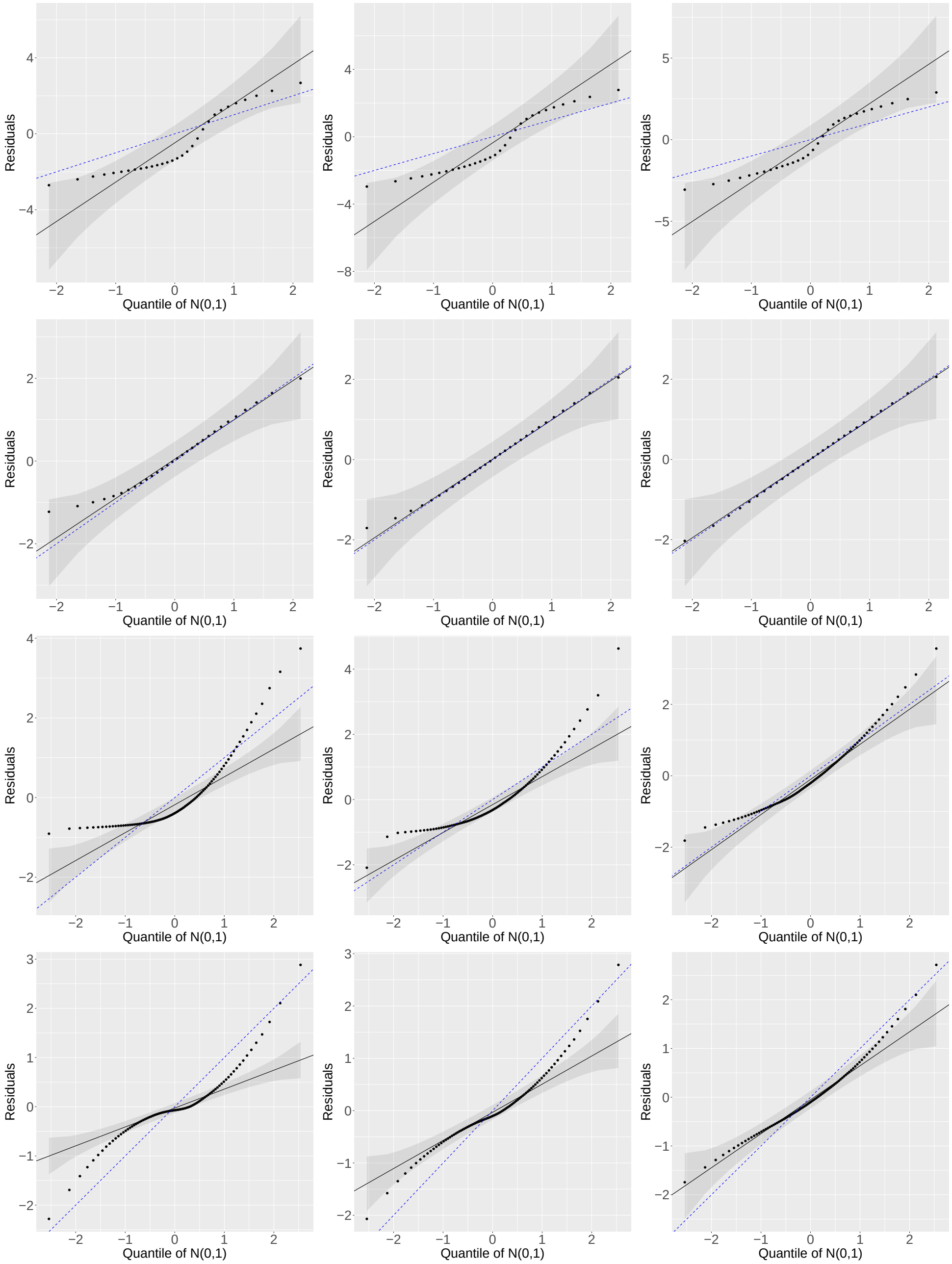

Figure 2. Normal probability plot of the standardized deviance component, randomized quantile, standardized Pearson and standardized weighted residuals (lines) for sample size $n = 30$, $\phi(\lambda) = 0.5(1.41),\ 1.0(1.0),\ 3.0(0.58)$ (columns) and $m_i = 3$ described in scenario 1.

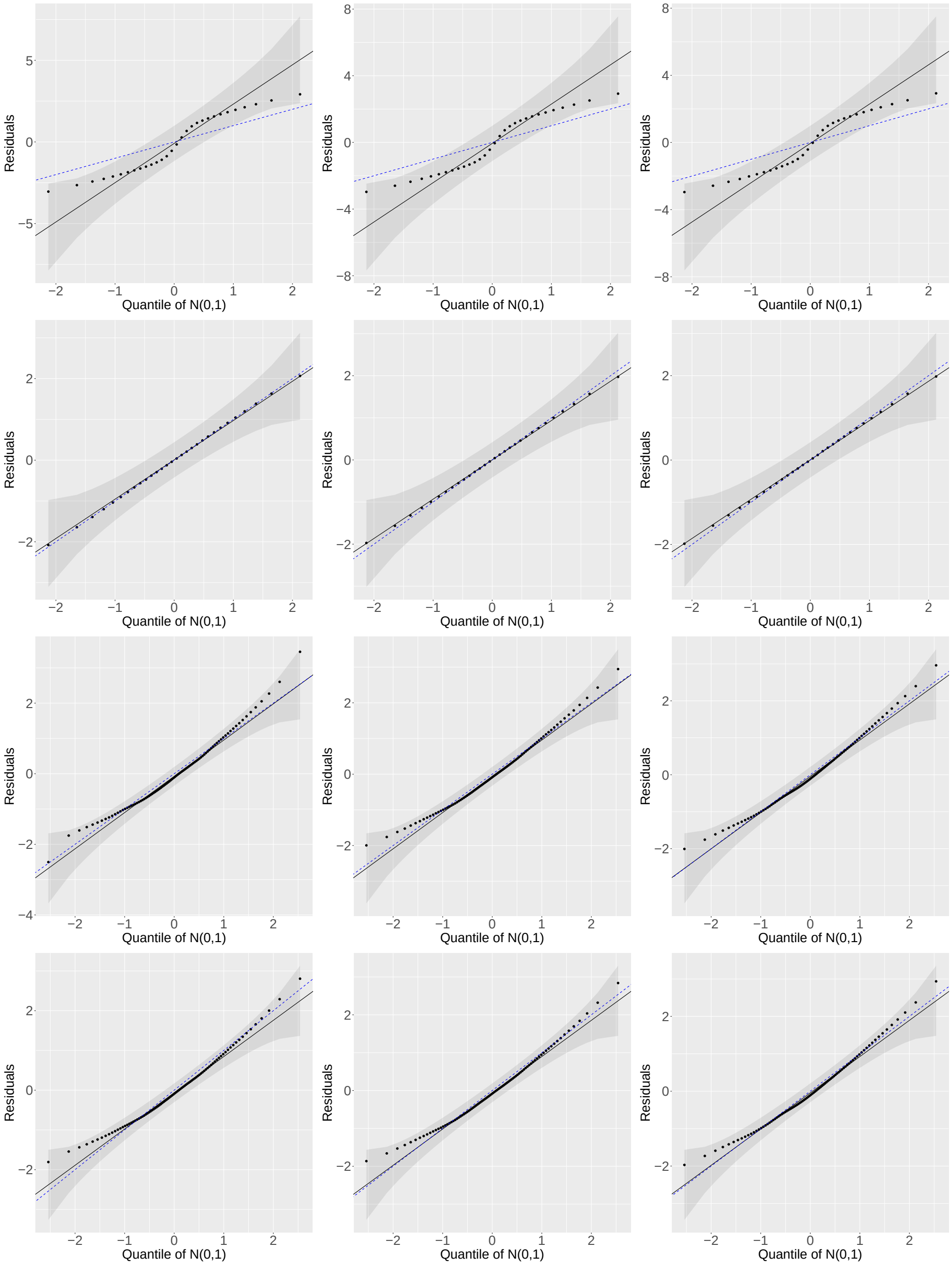

Figure 3. Normal probability plot of the standardized deviance component, randomized quantile, standardized Pearson and standardized weighted residuals (lines) for sample size $n = 30$, $\phi(\lambda) = 20(0.22),\ 100(0.10),\ 500(0.04)$ (columns) and $m_i = 3$ described in scenario 1.

Figures 2 and 3 show the normal probability plots of the order statistics mean of the standardized deviance component, randomized quantile, standardized Pearson and standardized weighted residuals. We consider confidence bands of 95% confidence level for the residual under study, as well as the fitted line (black line) and the 45-degree diagonal line (blue line). We expect that the residuals, as well as the black and blue lines, are superimposed. These figures reveal a high agreement between the quantile residuals and the standard normal distribution when the parameter $\phi > 1$; and when $\phi < 1$, the quantile residual present a little asymmetry. The empirical distribution of the standardized Pearson and standardized weighted residuals tends to be approximate to the standard normal distribution when the parameter ϕ reaches high values. Additionally, the empirical distribution of both residuals exhibits smooth tails to the right and to the left, featuring some asymmetry. Finally, the empirical distribution of the standardized deviance component residuals does not present a desirable behavior.

4.2 Scenario 2

In this second scenario, measures of central tendency are computed for the mean order statistics of the residuals defined in Section 3. Our purpose in this scenario is to study their asymmetric properties by considering three sample sizes, $n = 10, 100$, and 300. The results are displayed in the boxplot graphs in Figures 4 and 5. In the boxplot, the red point in the quantile interval corresponds to the sample mean and the blue lines correspond to the limits of the interval [-3,3].

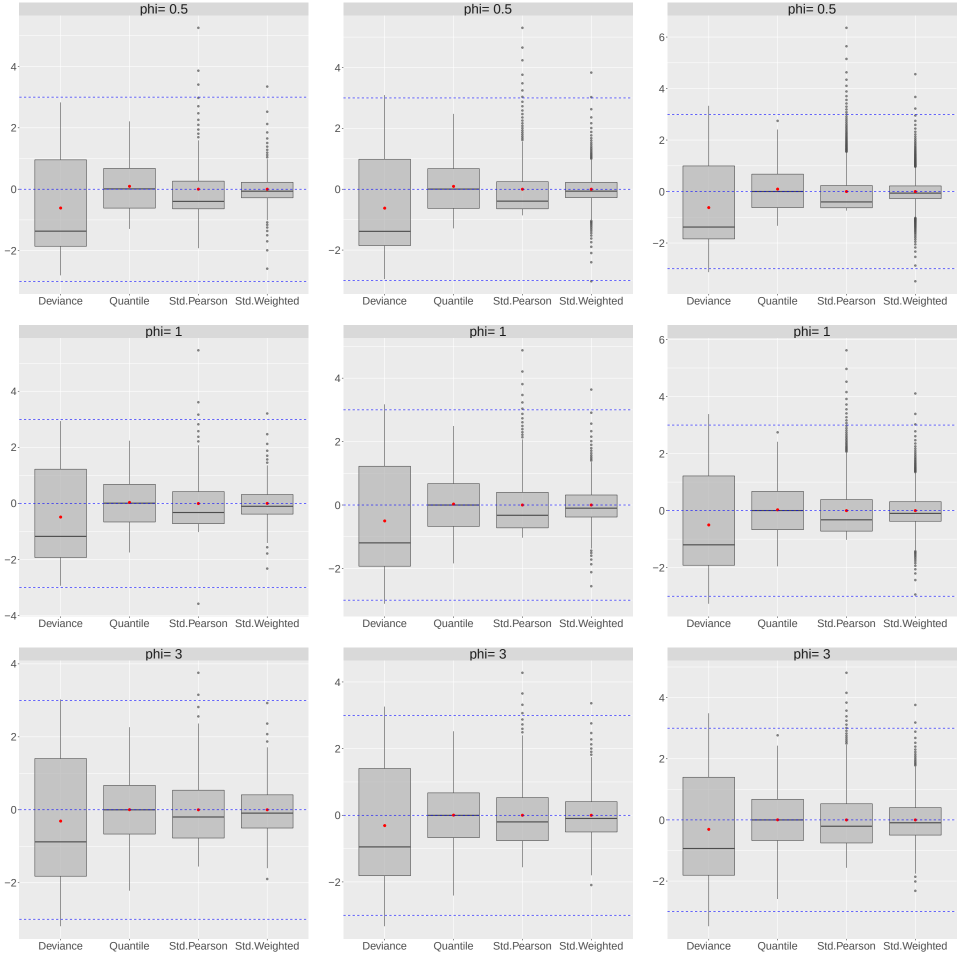

Figure 4. Boxplot of the standardized deviance component, randomized quantile, standardized Pearson and standardized weighted residuals for sample sizes $n = 10, 100, 300$ (columns), $\phi = 0.5, 1.0, 3.0$ and $m_i = 3$ described in scenario 2.

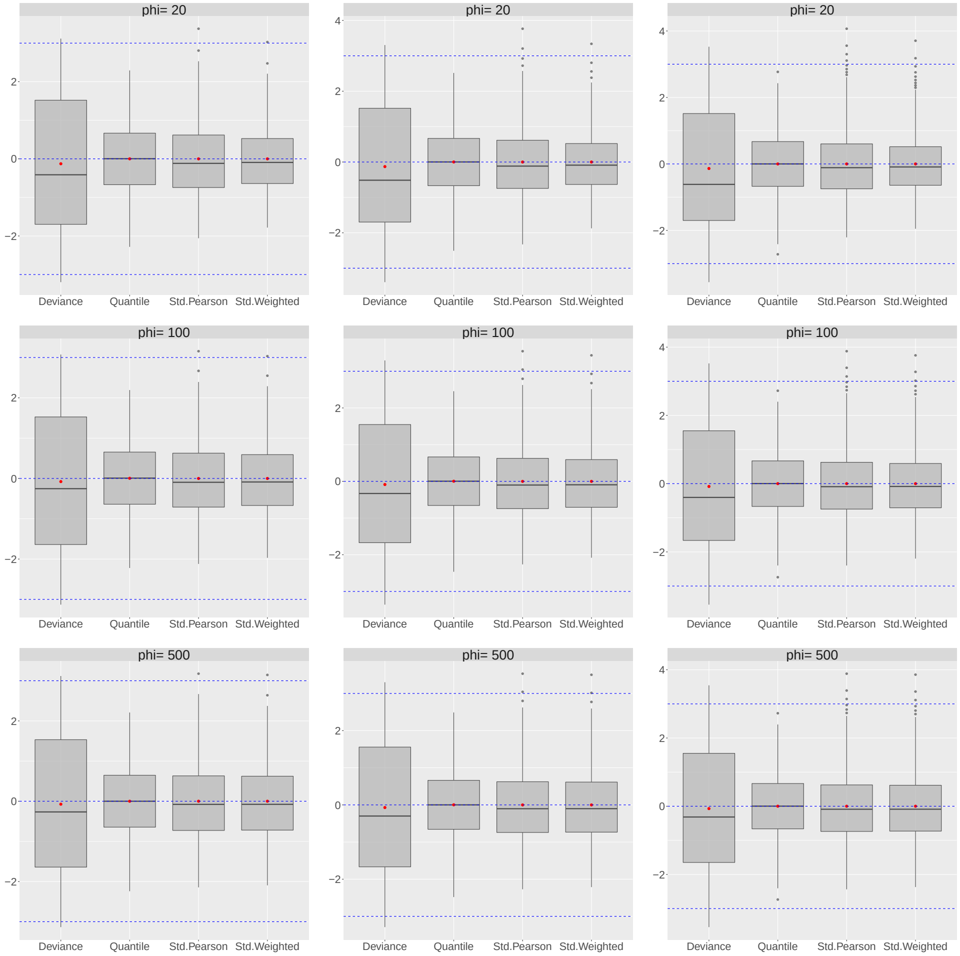

Figure 5. Boxplot graphics of the standardized deviance component, randomized quantile, standardized Pearson and standardized weighted residuals for sample sizes $n = 10, 100, 300$ (columns), $\phi = 20, 100, 500$ and $m_i = 3$ described in scenario 2.

Figures 4 and 5 reveal that, independent of the sample size, the quantile residuals preserve a symmetric behavior for the different values of ϕ. The standardized weighted residuals exhibit heavy tails when ϕ assumes values equal to 0.5 and 1.0 for all sample sizes. Its behavior becomes slightly asymmetric to the right for high values of $\phi > 3$, even though n increases. The standardized Pearson residuals are asymmetric to the right, decreasing when ϕ and n assume high values, respectively. Further, the simulation results for the standardized deviance component residuals only exhibit a nearly symmetric behavior when $n = 10$ and values of ϕ are high. According to these results, we can verify that the performance of these residuals is associated with the level of asymmetry of the shape parameter of the GLG distribution once the $\phi = \lambda^{-2}$. We also conclude that the randomized quantile residual is more appropriate for the cases in which $\phi < 1$, and the intraclass correlation tends to one. The standardized weighted residuals present better performance than Standardized Pearson. The quantile residuals present the desirable behavior.

4.3 Scenario 3

Figures 6 and 7 show the empirical distribution of the residuals compared with the quantiles of the standard normal distribution and the boxplot graphs, respectively, when the random effect distribution is misspecified. Even though we assumed a normal distribution for the random effect distribution and fitted by MNBR, the random quantile residuals show a high agreement with the standard normal distribution for all ϕ and n values. The standardized Pearson weighted residuals have this behavior for high values of ϕ (see Figure 6). However, the boxplot graphs reveal that the standardized weighted residuals exhibit a smooth asymmetry to the right for values of $\phi \geq 3$, in agreement with the simulation results present in Scenario 2. Based on all simulation results, we conclude that the standardized weighted residuals can be an alternative to quantile residuals as a goodness of fit for the MNBR model when $\phi \geq 3$.

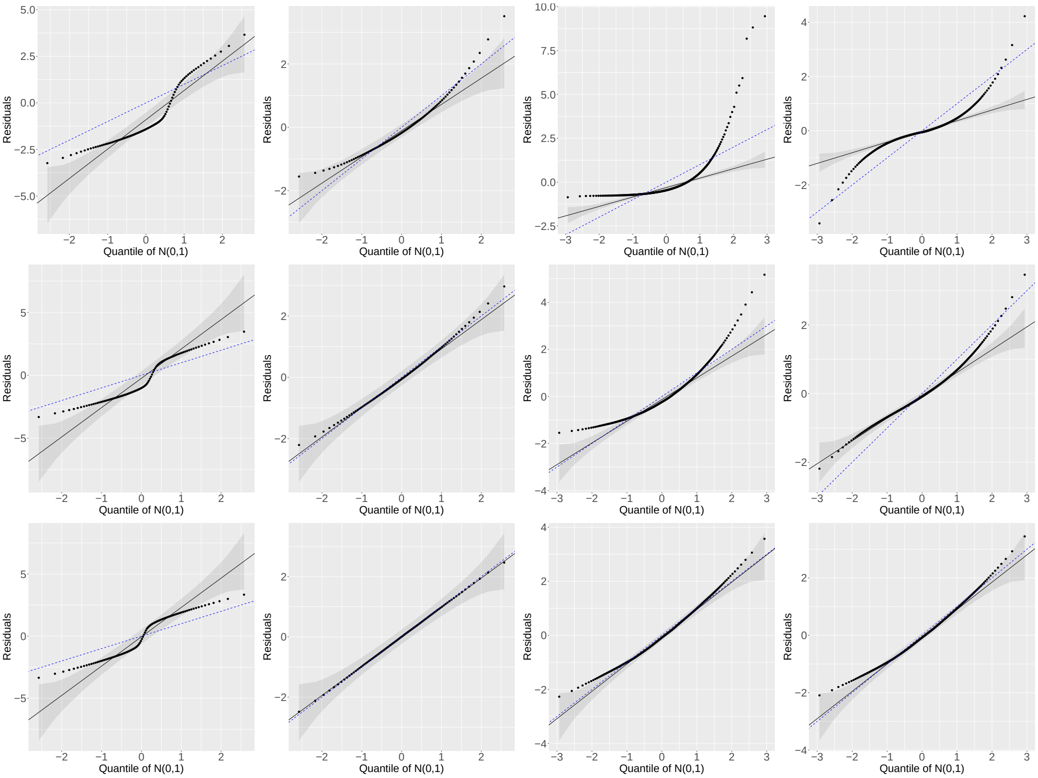

Figure 6. Normal probability plot of the standardized deviance component, randomized quantile, standardized Pearson and standardized weighted residuals (columns) for sample size $n = 100$, $\phi = 0.5,\ 3.0,\ 100$ (lines) and $m_i = 3$ described in scenario 3.

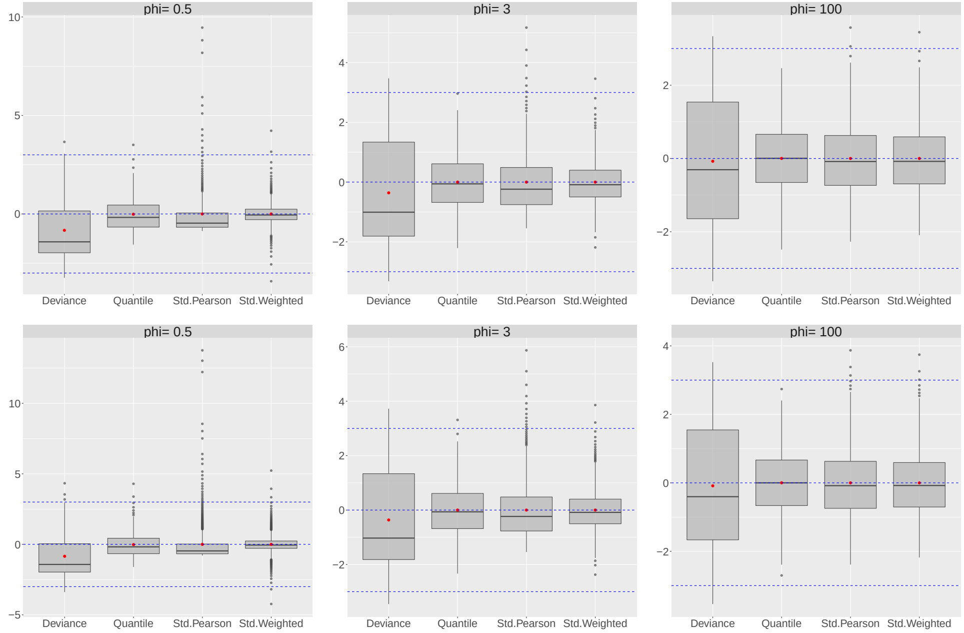

Figure 7. Boxplot of the standardized deviance component, randomized quantile, standardized Pearson and standardized weighted residuals for sample sizes $n = 100, 200$, (lines) $\phi = 0.5, 3.0, 100$ (columns) and $m_i = 3$ described in scenario 3.

In addition, we conducted an additional simulation study (it does not show the results) to evaluate the behavior of the empirical distribution of the residuals with a small means. We considered a unique sample size of $n = 10$, the real values to the vector of parameters as $(\phi, \beta_0, \beta_1)^\top = (3.0, 0.5, -0.5)^\top$ and three different sub scenarios. The random samples are generate assuming: (*i*) $x_{ij1} \sim U(-1.5; 1.5)$ (true model), (*ii*) $x^*_{ij1} = x^2_{ij1}$ (wrong model) and (*iii*) $x^{**}_{ij1} = 1/x_{ij1}$ (wrong model). The maximum estimates are obtained assuming the linear systematic model, i.e., $g(\mu_{ij}) = \beta_0 + \beta_1 \times x_{ij1}$. Unlike Feng *et al.* (2020), we observed that the residuals, when the covariate is misspecificated, have a very close empirical distribution when the true model is assumed. Also, the empirical distributions of the residual have the same behavior as previous simulation results.

5. Applications

5.1 Alzheimer's data

The Alzheimer's data is presented in Hand & Taylor (1987) and Hand & Crowder (1996) assess the deterioration aspects of intellect, self-care, and personality in senile patients with Alzheimer's disease. Two groups of patients were compared, one of which received a placebo and the other received treatment with Lecithin. In the data, each of the subjects, 26 in the placebo group and 22 in the Lecithin group, were measured on five occasions (initially, 1st, 2nd, 4th, and 6th). The measurements were the number of words that the patients could recall from a list of words. The major interest in this study is to investigate whether the memory-effect differs between the two treatment groups. Figure 8, exhibits the presence of atypical patients in the placebo and the Lecithin groups, measured at different times. The mean number of recorded words in each measurement is indicated by the black point. Its behavior provides evidence that the mean number of words recorded for the placebo group does not differ from that of the Lecithin group. This fact is evidenced in Table 1, as well as the smaller dispersion in the Lecithin group. Further, the rate between the variance and the mean in each record is an indicator of the overdispersion phenomenon present in the data. Based on this fact, we propose the MNBR model for fitting the data and accommodating the extra variability.

Table 1. Summary of Placebo and Lecithin group for the Alzheimer's data

	Placebo				
	1	2	3	4	5
Mean	9.65	9.23	10.08	8.92	8.73
Variance	26.87	29.78	26.07	31.19	27.56
Variance/Mean	2.78	3.22	2.58	3.49	3.15
	Lecithin				
	1	2	3	4	5
Mean	8.14	9.77	8.41	9.45	9.05
Variance	22.50	21.89	21.39	29.59	26.62
Variance/Mean	2.76	2.24	2.54	3.13	2.94

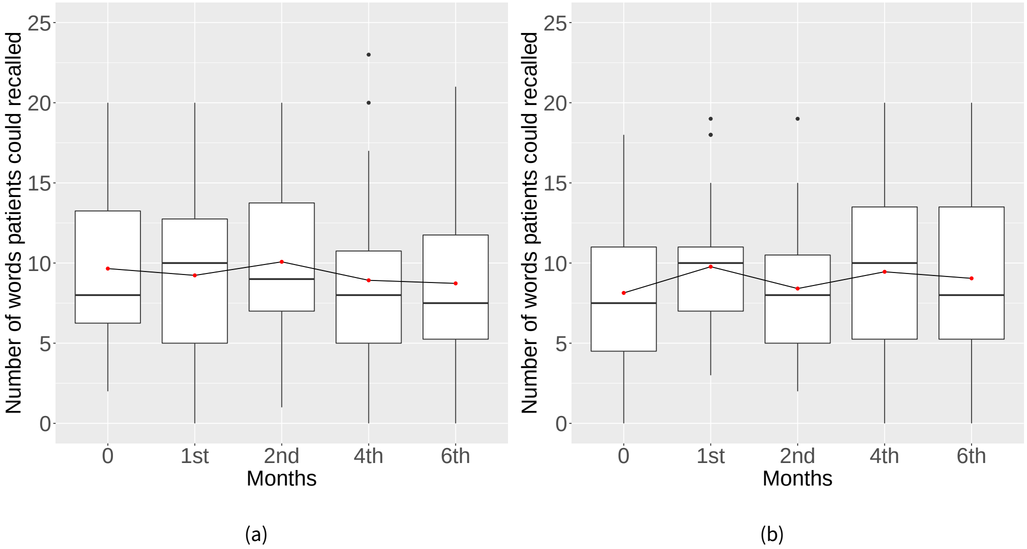

Figure 8. Boxplot of the patients in (a) the placebo and (b) the lecithin group, respectively for the Alzheimer's data.

Thus, we consider that the vector response containing the number of words recalled by the i-th patient follows the hierarchical structure: (i) $\boldsymbol{y}_i \overset{\text{ind}}{\sim} \text{MNB}(\boldsymbol{\mu}_i, \phi)$ and (ii) $\log(\boldsymbol{\mu}_i) = \beta_1 + \beta_2 \text{Group}_i$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i5})^\top$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{i5})^\top$, for $i = 1, \ldots, 48$, β_1 intercept and β_2 represent the Lecithin effect with respect to the placebo. The parameter estimates obtained using the are shown in Table 2. The inference results reveal that the mean number of words recalled in the Lecithin and placebo groups is the same statistically. The standardized weighted, standardized Pearson, and quantile residuals are employed for a residual analysis of the MNBR model fitted to Alzheimer's data. Figure 9 exhibits two graphs. In the first line, the plots of these residuals against the observations show that they are randomly spread around zero between the threshold of ±3 without the presence of outliers. Their respective normal plots with simulation envelopes indicate the adequacy of the MNBR model to the Alzheimer's data. The performance of these three residuals is in accordance with the simulation study when ϕ assumes a value close to three.

Table 2. Parameter estimates, standard errors (Std. error), z-values, and p-values for the MNBR model fitted to the Alzheimer's data

Parameter	Estimate	Std. error	z-value	p-value
ϕ	3.811	0.834	–	–
β_1	2.232	0.104	21.367	0.000
β_2	-0.039	0.154	-0.254	0.799

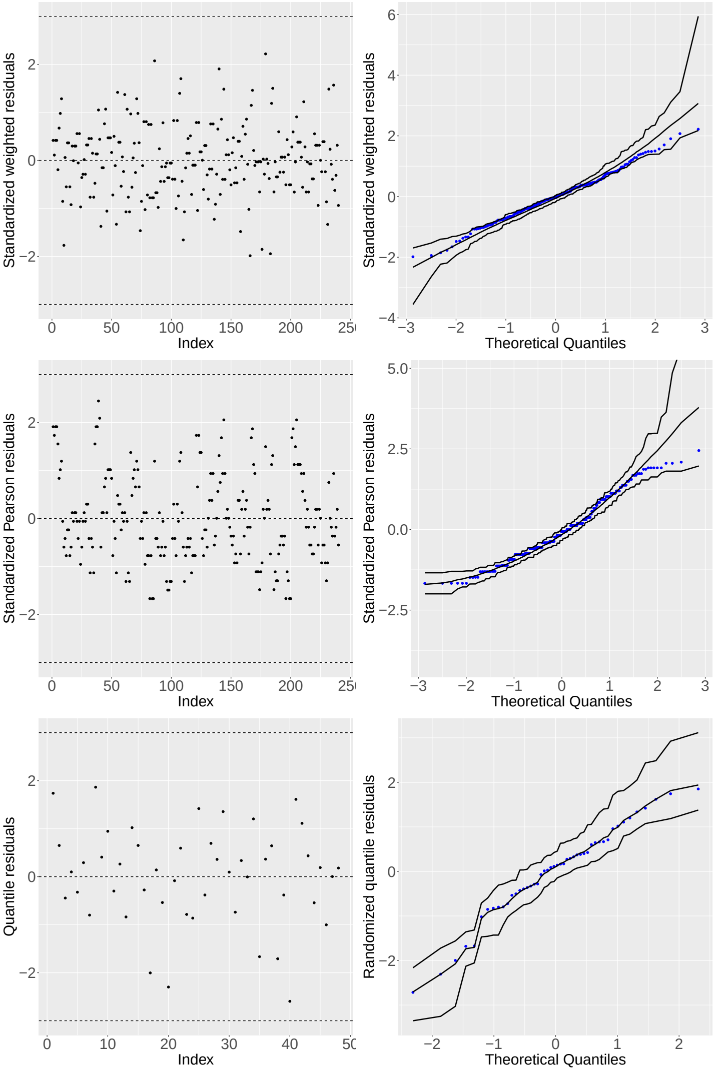

Figure 9. The residuals and simulated envelope plots of the standardized weighted, standardized Pearson and randomized quantile residuals to the Alzheimer's data.

5.2 Seizure's data

The data set described in Diggle *et al.* (2013) and recently by Fabio *et al.* (2023) refers to an experiment in which 59 epileptic patients were randomly assigned to one of the two groups the treatment (Progabide drug) and placebo groups. The number of seizures experienced by each patient during the baseline period (week eight) and the four consecutive periods (every two weeks) was recorded. The main goal of this application is to analyze the drug effect with respect to the placebo. Two dummy covariates are considered in this study: Group, which assumes values equal to 1 if the patient belongs to the treatment group and 0 otherwise; and Period, which assumes values equal to 1 if the number of seizures is recorded during the treatment and 0 if they are measured in the baseline period. Considering the irregular measurement of rate of seizures during the time, the variable Time is considered as an offset for fitting the data, where Time assumes value equal to 8 if the number of seizures is observed in the baseline period and 2 otherwise.

In the boxplot graph in Figure 10, we note atypical individuals, patients $\sharp 18(111, 37, 29, 28, 29)$ and $\sharp 25(55, 18, 24, 76, 25)$ that belong to the placebo group. These, atypical observations present a high number of seizures in the baseline period and in the third visit of the $\sharp 25$ patient compared to other clinic visits. Patients $\sharp 29(76, 11, 14, 9, 8)$ and $\sharp 49(151, 102, 65, 72, 63)$ belong to the Progabide group, and they experienced a decrease in the number of seizures in which clinic visit. This indicates the effectiveness of the drug in patients with complex seizures. The average number of seizures in each record is indicated by the red point. Its behavior is similar in both the groups, indicating that the number of seizures over time does not differ between the groups. Fabio *et al.* (2023) showed the occurrence of the overdispersion phenomenon in the seizure data.

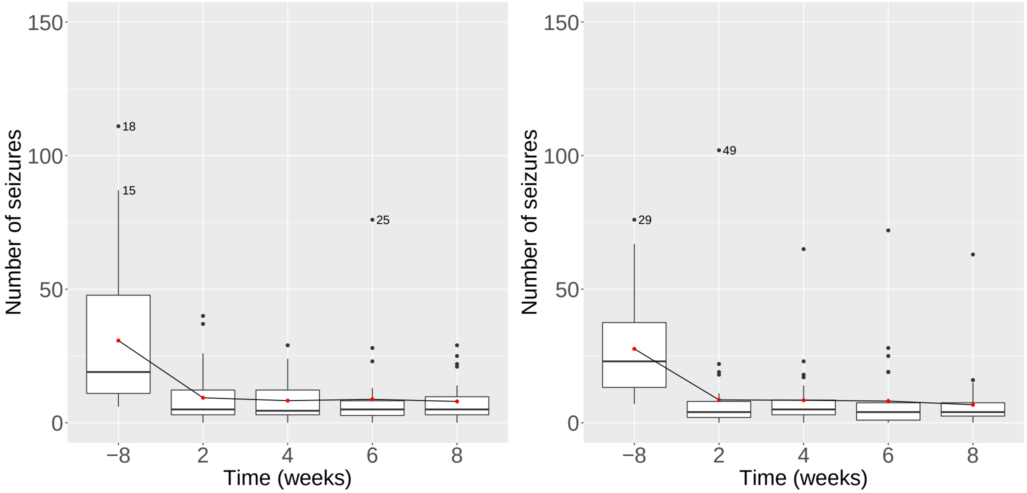

Figure 10. Boxplot and profile mean for placebo (left) and treatment (right) group for the seizure's data.

The parameters estimated are obtained assuming the following structure: (*i*) $\boldsymbol{y}_i \overset{\text{ind}}{\sim} \text{MNB}(\boldsymbol{\mu}_i, \phi)$ and (*ii*) $\log(\mu_{ij}) = \beta_1 + \beta_2\text{Group}_i + \beta_3\text{Period}_{ij} + \beta_4(\text{Group}_i \times \text{Period}_{ij}) + \log(\text{Time}_{ij})$, where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})^\top$, $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4})^\top$, for $i = 1, \ldots, 59$, β_2 is the logarithm of the ratio of the average rate of the treatment group to the placebo group at baseline, β_3 is the logarithm of the ratio of the seizure mean after the treatment period to before the treatment period for the placebo group, and $\exp(\beta_4)$ is the treatment effect, and it is the ratio of post- to pre-treatment mean seizure ratios between the treatment and placebo groups. The parameter estimates obtained are shown in Table 3. The estimates confirm that there is not enough evidence of the treatment effect.

Table 3. Parameter estimates with their respective approximate standard errors (Std. error), z-values, and p-values for the MNBR model fitted for the seizure's data

Parameter	Estimate	Std. error	z-value	p-value
ϕ	1.607	0.278	–	–
β_1	1.348	0.153	8.813	< 0.001
β_2	0.028	0.211	0.131	0.896
β_3	0.112	0.047	2.386	0.017
β_4	-0.105	0.065	-1.610	0.107
λ	0.789			

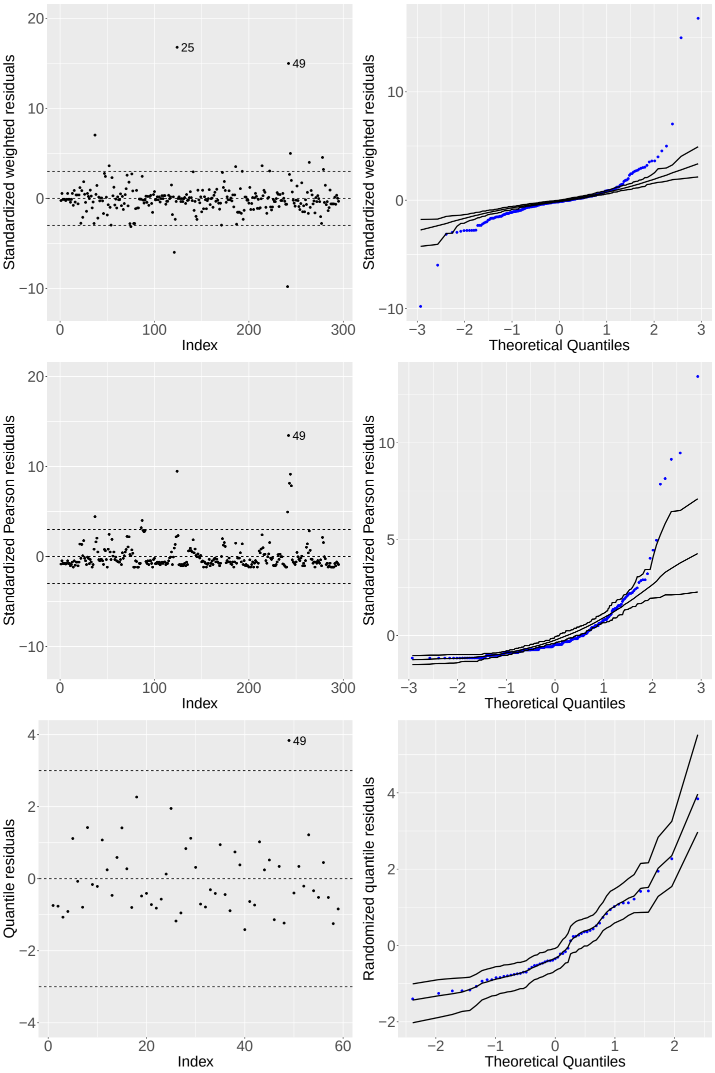

Figure 11. The residuals and simulated envelope plots of the standardized weighted, standardized Pearson and randomized quantile residuals to the seizure's data.

The standardized weighted, standardized Pearson, and quantile residuals are employed for a residual analysis of the MNBR model fitted to seizure data. Figure 11 suggested that the quantile residual is more appropriate for checking the departures of the MNBR model fitted to seizure data. Simulation studies show that standardized Person and standardized weighted residuals are inappropriate when ϕ assumes small values (In our application, $\phi = 1.607$).

6. Conclusion

The empirical distribution of the four residuals is assessed for the MNBR model. We proposed the standardized weighted and standardized Pearson residuals, and, to complement our study, we considered the standardized component of deviance and quantile residuals suggested by Fabio *et al.* (2012) and Fabio *et al.* (2023), respectively. Monte Carlo simulation studies were carried out to evaluate the approximation of the empirical distributions of the residuals with respect to the standard normal distribution. We verify that the performance of these residuals depends on the dispersion parameter, which is associated with the level of asymmetry of the shape parameter of the GLG distribution, once $\phi = \lambda^{-2}$. We conclude that the quantile residual is suggested when $\phi > 0.5$, for all sample sizes. The standardized weighted residuals can be an alternative to quantile residuals when $\phi \geq 3$ for every n. The standardized Pearson residual can be employed as the parameter ϕ and n assume high values. Finally, we suggest drawing the normal probability plots with simulation envelope for standardized Pearson and weight residuals to check the adequacy of the MNBR model to the data by considering the simulation results.

Acknowledgments

The authors thank the reviewers for valuable suggestions that improved the paper. The last author gratefully acknowledges the funding provided by Programa Institucional de Internacionalização – CAPES – PrInt – (Grant No. 88887.935790/2024-00).

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: FABIO L.C.; CARRASCO, J.M.F. **Data curation**: FABIO L.C. **Formal analysis**: FABIO L.C.; VILLEGAS, C. **Funding acquisition**: CARRASCO, J.M.F. **Investigation**: FABIO L.C.; CARRASCO, J.M.F. **Methodology**: FABIO L.C.; CARRASCO, J.M.F. **Project administration**: FABIO L.C.; CARRASCO, J.M.F. **Software**: CARRASCO, J.M.F.; VILLEGAS, C. **Resources**: FABIO L.C.; CARRASCO, J.M.F.; MAMUN, A.S.M.A. **Supervision**: FABIO L.C.; CARRASCO, J.M.F. **Validation**: FABIO L.C.; CARRASCO, J.M.F.; MAMUN, A.S.M.A. **Visualization**: FABIO L.C.; CARRASCO, J.M.F. **Writing - original draft**: FABIO L.C.; CARRASCO, J.M.F. **Writing - review and editing:** FABIO L.C.; CARRASCO, J.M.F.; VILLEGAS, C.; MAMUN, A.S.M.A.

References

1. Agresti, A. *Foundations of Linear and Generalized Linear Models* (Wiley, New Jersey, 2015).
2. Diggle, P. J., Liang, K. Y. & Zeger, S. L. *Analysis of Longitudinal Data* 2nd ed. (Oxford University Press, N.Y., 2013).
3. Espinheira, P. L., S., F. & Cribari-Neto, F. On beta regression residuals. *Journal of Applied Statistics* **35,** 407–419. https://doi.org/10.1080/02664760701834931 (2008).
4. Fabio, L. C., Paula, G. A. & de Castro, M. A Poisson mixed model with nonormal random effect distribution. *Computational Statistics and Data Analysis* **56,** 1499–1510. https://doi.org/10.1016/j.csda.2011.12.002 (2012).
5. Fabio, L. C., Villegas, C., Carrasco, J. M. F. & de Castro, M. Diagnostic tools for a multivariate negative binomial model for fitting correlated data with overdispersion. *Communications in Statistics - Theory and Methods.* **52,** 1833–1853. https://doi.org/10.1080/03610926.2021.1939380 (2023).
6. Faraway, F. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and nonparametric regression models* (Chapman and Hall, New York, 2016).
7. Feng C., L., L. & Sadeghpour, A. A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology* **20,** 1–21. https://doi.org/10.1186/s12874-020-01055-2 (2020).

8. Hand, D. J. & Crowder, M. *Practical Longitudinal Data Analysis* (London: Chapman & Hall, 1996).

9. Hand, D. J. & Taylor, C. C. *Analysis of Variance and Repeated Measures* (London: Chapman & Hall, 1987).

10. Hardin, J. W. & Hilbe, J. M. *Generalized Linear Models and Extensions, Second Edition* (Stata Press, Texas, 2016).

11. Johnson, N., Kotz, S. & Balakrishnan, N. *Discrete Multivariate Distributions* (Wiley, New York, 1997).

12. Lawless, J. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15,** 209–225. https://doi.org/10.2307/3314912 (1987).

13. Pereira, G. H. A., Scudilio, J., Santos-Neto, M., Botter, D. A. & Sandoval, M. C. A class of residuals for outlier identification in zero adjusted regression models.*Journal of Applied Statistics* **47,** 1833–1847. https://doi.org/10.1080/02664763.2019.1696759 (2020).

14. Scudilio, J. & Pereira, G. H. A. Adjusted quantile residual for generalized linear models. *Computational Statistics* **35,** 399–421. https://doi.org/10.1007/s00180-019-00896-w (2020).

15. Tsui, K.-W. Multiparameter estimation for some multivariate discrete distributions with possibly dependent components. *Annals of the Institute of Statistical Mathematics* **38,** 45–56. https://doi.org/10.1007/BF02482499 (1986).

16. Waller, L. A. & Zelterman, D. Log-Linear Modeling with the Negative Multinomial Distribution. *Biometrics* **53,** 971–982. https://doi.org/10.2307/2533557 (1997).