Jussiane Gonçalves da Silva

Zero-Inflated Mixed Poisson Regression Models

Belo Horizonte

2017

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Zero-Inflated Mixed Poisson Regression Models

Jussiane Gonçalves da Silva
Author

Wagner Barreto de Souza, PhD.
Supervisor

Belo Horizonte
2017

Acknowledgments

First of all, I would like to thank God that has given me the strength to overcome the obstacles in my way. I would like to thank my family, Iracema, Varley, Tassiane, Eloy and Mirela, those who have supported me in this journey. I also appreciated my colleagues' help from UFMG, especially those from 3036 lab. I would like to especially thank professor Wagner, who has taught me and has mentoring me over the past two years, sharing his knowledge and being patient. I would also like to thank CAPES, since this work was accomplished with their financial support.

Let us hold unswervingly to the hope we profess,

for He who promised is faithful.

Hebrews 10:23

Resumo

O modelo de regressão de Poisson é muito utilizado para o ajuste de dados de contagem, conforme esclarecem Lawless (1987) and Karlis (2001), porém um problema que surge da utilização dessa distribuição é que ela é equidispersa, ou seja, a variância é igual à média. Porém, o que se observa na prática é que os dados são sobredispersados em sua grande maioria. Na literatura é possível encontrar diversos modelos que lidam com o problema da sobredispersão, como é o caso dos modelos de regressão binomial negativa (NB) [Lawless (1987)] e Poisson-inversa Gaussiana (PIG) [Dean et al. (1989)], derivados de misturas de Poisson. Uma classe geral de modelos de regressão de Poisson misturada foi introduzida por Barreto-Souza and Simas (2016).

Contudo, o excesso de zeros em dados de contagem é um fator que leva à sobredispersão e, quando a taxa de inflação de zeros é muito elevada, os modelos de mistura de Poisson não são suficientes para adequar a variabilidade, conforme explicam Dean and Nielsen (2007). Então, para contornar esse problema, modelos zero-inflados são facilmente encontrados na literatura, como é o caso dos modelos de regressão Poisson zero-inflado, introduzido por Lambert (1992), binomial negativa zero-inflado, utilizado por Yau et al. (2003), e Poisson generalizada zero-inflado, introduzida por Famoye and Singh (2006). Além disso, dados zero-inflacionados são encontrados em diversas áreas, como biologia [Oliveira et al. (2016)], manufatura e engenharia [Lambert (1992), Li et al. (1999)], agricultura [Ridout et al. (2001)], saúde [Mwalili et al. (2008), Lim et al. (2014)], ciências sociais [Famoye and Singh (2006)], entre outras.

Dessa forma, o objetivo deste trabalho é fornecer suporte apropriado para lidar com dados de contagem sobredispersados e o excesso de zeros e, para tal, propõe-se um modelo de regressão geral com base numa classe de distribuições de misturas de Poisson zero-infladas, onde unifica-se modelos já consolidados, como os modelos ZINB e ZIPIG, bem como permite o surgimento de novos modelos zero-inflados.

Portanto, está sendo proposto uma classe geral de modelos de regressão de Poisson misturada zero-inflado para lidar, simultaneamente, com a sobredispersão e o excesso de zeros. Logo, em relação aos recursos computacionais, propôs-se obter as estimativas dos parâmetros do modelo por meio do algoritmo EM, que consegue lidar com a estrutura latente existente. Além disso, são fornecidas as expressões explícitas para obtenção da matriz de informação sendo possível, dessa forma, obter os desvios padrão das estimativas dos parâmetros, o que permite, por exemplo, a construção de intervalos de confiança.

Um estudo de simulação foi executado para avaliar o comportamento das estimativas obtidas por meio do algoritmo EM, como por exemplo o comportamento para amostras de tamanho pequeno, bem como também avaliar a matriz de informação estimada. Ademais, para investigar pontos discrepantes e sua possível influência, uma análise de resíduos foi executada, com base na simulação de envelopes. Com o objetivo de aferir a influência global de outliers, está sendo utilizada a distância de Cook generalizada, proposta por Zhu et al. (2001), tendo sido fornecidas as expressões explícitas dessa medida para o modelo proposto, objetivando assim checar a adequabilidade da distribuição assumida para a variável resposta.

Palavras-chave: excesso de zeros, modelos de regressão para contagens, sobredispersão, algoritmo EM.

Abstract

When someone is dealing with discrete response variables, the Poisson regression model is commonly used for fitting count data, just as described by Lawless (1987), Karlis (2001) and Sellers and Shmueli (2010), for instance. However, a drawback of this model is that Poisson distribution is equidispersed, that is, variance equal to mean. In practice, overdispersed count data are often observed and to handle this problem, different kind of mixed Poisson regression models, wherein the variance is larger than the mean, have been introduced in the literature, such as the negative binomial (NB) regression model [Lawless (1987)], which is widely used, and Poisson-inverse Gaussian (PIG) regression model [Dean et al. (1989), Holla (1967)]. A general class of mixed Poisson regression models was proposed by Barreto-Souza and Simas (2016).

As mentioned by Garay et al. (2011) and Barreto-Souza and Simas (2016), a factor that can lead to overdispersion is the excess of zeros in count data, although Dean and Nielsen (2007) clarify that mixed Poisson regression models may not be suitable for modeling data with high zero-inflation rate, since overdispersion may remain. To deal with this issue, zero-inflated models for count data are easily found in literature, such as zero-inflated Poisson (ZIP) regression model, introduced by Lambert (1992), zero-inflated negative binomial (ZINB) regression model [see, for example, Yau et al. (2003)] and, moreover, zero-inflated generalized Poisson (ZIGP) regression models, proposed by Famoye and Singh (2006).

Data with too many zeros are easily encountered in several fields, as biology [Oliveira et al. (2016)], manufacturing application and engineering [Lambert (1992), Li et al. (1999)], agriculture [Ridout et al. (2001)], health [Mwalili et al. (2008), Lim et al. (2014)], social sciences [Famoye and Singh (2006)] and many other disciplines. Thus, in order to deal with overdispersion and the excess of zero in count data, the goal of this work is to provide appropriated support for modeling these kind of data. For this purpose, we are proposing a general regression model based on a class of zero-inflated mixed Poisson distributions. With this approach, we unify some existent models, such as ZINB and zero-inflated Poisson-inverse Gaussian (ZIPIG) regression models, as well as open the possibility of introducing new zero-inflated models.

Keywords: zero-inflation, count regression models, overdispersion, EM algorithm.

List of Figures

List of Tables

List of Abbreviations

BFGS ..... Broyden-Fletcher-Goldfarb-Shanno Algorithm

RSDS ..... Root of Square Difference Sum

EF ..... Exponential Family

EM ..... Expectation-Maximization Algorithm

GAMLSS .... Generalized Additive Models for Location, Scale and Shape

GCD ..... Generalized Cook's Distance

GPR ..... Generalized Poisson Regression Model

MLE ..... Maximum Likelihood Estimates

MP ..... Mixed Poisson

NB ..... Negative Binomial

PIG ..... Poisson-inverse Gaussian

RMSE ..... Root of the Mean Square Error

ZIB ..... Zero-inflated Binomial

ZIGP ..... Zero-inflated Generalized Poisson

ZIMP ..... Zero-inflated Mixed Poisson

ZINB ..... Zero-inflated Negative Binomial

ZIP ..... Zero-inflated Poisson

ZIPIG ..... Zero-inflated Poisson-inverse Gaussian

Contents

1 Introduction

The regression models that handle data with zeros excess arises at a first moment from the model proposed by Lambert (1992), the zero-inflated Poisson (ZIP) regression model. The model is a mixture of the Poisson distribution and a point mass of one at zero and, with this, the count of zeros may come from two sources, it is, from the Poisson distribution (sampling zeros) or from the point of mass, named as structural zeros. Lambert (1992) gives motivation for the ZIP model explaining that, in manufacturing process, the equipment is near to no defect when it is properly aligned and it is the source of the structural zero defects. But when the equipment is misaligned, then the defects number of the equipment may come from a Poisson distribution. The estimates of the model were obtained through the EM algorithm proposed by the author, maximizing iteratively the incomplete log-likelihood function. According to the author, the EM algorithm converges and reasonably fast.

Despite the ZIP models without covariates have already been studied earlier by Cohen (1963), Johnson and Kotz (1969), the author explains that the parameter related to the structural zeros and the mean of the Poisson distribution may depend on covariates. This argument is strengthened by Cameron and Trivedi (1998), that explain that regression analysis of counts are motivated by the observation that, in many real situations, the assumption that samples are independent and identically distributed is too strong. They complete their argument giving an example, explaining that the occurrence of an event may depend on some covariates and can vary from case to case, accomplishing by a regression model for event count.

In the paper of Ridout et al. (1998), the authors speech about the problems of the zeros excess and review some models that can handle these issues. They clarify that despite the Poisson regression model provides a standard structure for modeling count data, count data frequently are overdispersed relative to the Poisson distribution, what

turns the Poisson models inappropriate. The authors also explain that the incidence of zeros is a source of overdispersion because it is greater than the expected from the Poisson distribution. According to Ridout et al. (1998), the interest in models that deal with zero-inflation has been substantial and a reason is because zero counts have a special status, it is, the structural zeros and the ones that come from de count distribution, as mentioned by them.

In their models review, Ridout et al. (1998) start talking about the mixed Poisson distributions, that have been oftentimes used for modeling overdispersed data. The overcome can be noticed considering that a random variable Y follows the Poisson distribution with μV parameter, where V is a random variable with one as expected value and α as variance. Then, $E(Y) = \mu$ and $\text{Var}(Y) = \mu + \alpha\mu^2$, which is larger than the mean.

Subsequently they introduce what they call as zero-modified distributions, presenting first the ZIP distribution given by

$$P(Y = y) = \begin{cases} \omega + (1-\omega)\exp(-\lambda), & y = 0 \\ (1-\omega)\exp(-\lambda)\lambda^y/y!, & y > 0 \end{cases},$$

where Y follows the ZIP distribution, ω is the structural zeros parameter and λ is the Poisson distribution parameter. According to the authors, a zero-deflated model can be derived considering ω as a negative number, although the distribution cannot arise from a mixture and, moreover, zero-deflated data are uncommon in practice. Then, they explain that other models can arise with the same structure presented at the expression above, for example when one uses the negative binomial or the generalized Poisson distributions instead of the Poisson distribution.

Thereafter, the authors present other kinds of models to deal with the excess of zeros, such as the hurdle models introduced by Mullahy (1986), it is, a two part models, wherein the first one is ruled by the binomial probability model which determines whether a zero or non-zero outcome occurs, while the second part is ruled by a truncated count distribution. The idea is that given an event has occurred, that is, the "hurdle has been crossed", the conditional distribution of this event is controlled by a truncated at zero distribution. Finely, they present some inferential aspects, as some criterion for comparison between models and, moreover, they fitted the Poisson, NB, ZIP and the ZINB regression models for a horticultural count data, the same that has been used in the empirical illustration of this work.

Applications of the ZINB regression model are easily encountered in literature, especially because the ZIP regression model may not be adequate for some data in which there is evidence of overdispersion after the fit, it is, when the overdispersion remain. According to Famoye and Singh (2006), the ZINB regression model is not always a good choice, since the model could not be fitted to some data sets because its failure on the convergence by a iterative technique for parameter estimation. A similar observation was pointed by Lambert (1992), explaining that the ZINB regression model is possibly a better model for the manufacturing case, however the ZINB regression model did not succeed in fitting the data set. That issue was the motivation for Famoye and Singh (2006) develop the ZIGP regression model in order to fit overdispersed count data with many zeros. The probability function of the zero-inflated generalized Poisson distribution is given by

$$P(Y=y)=\begin{cases}\varphi+(1-\varphi)f(\mu,\alpha;y), & y=0\\ (1-\varphi)f(\mu,\alpha;y), & y>0\end{cases},$$

where Y follows the ZIGP distribution and $f(\mu, \alpha; y)$ is the generalized Poisson probability function, given by

$$f(\mu, \alpha; y) = \left(\frac{\mu}{1+\alpha\mu}\right)^y \frac{(1+\alpha y)^{y-1}}{y!} \exp\left[\frac{-\mu(1+\alpha y)}{1+\alpha\mu}\right], \quad y = 0, 1, 2, \ldots.$$

An interesting point of the ZIGP model is that it allows the fit of a zero-deflated model by using some appropriate link function that may enables φ assumes negative values, but the authors pay attention that zero-deflation cases seldom occurs in practice. A second remark is that the ZIGP model reduces to the ZIP model when $\alpha = 0$.

The authors also present a score test for the model and highlight the importance of the test explaining that one maybe does not need to fit the ZIPG regression model, but just the generalized Poisson regression model (GPR), which is the distribution under the null hypothesis. Then, the score statistic, that has an asymptotic chi-square distribution with one degree of freedom, will reveal if the GPR model fits well the number of zeros.

An application is presented in the article, using a domestic violence data set, where the authors argue, throughout the score test, that there are too many zeros in the data and through a measure of goodness-of-fit they conclude that ZIGP regression model is more adequate than ZIP regression model, since the α parameter that reduces ZIGP model to the ZIP model is significant different of zero. The conclusion is that ZIGP regression model fits well the domestic violence data and its also a competitor to the ZINB regression model. However, the authors do not know in which terms one can be better than the another, it is, which is the better model, they only point that in few cases the ZINB regression model did not converges.

Another example of zero-inflated model is the zero-inflated Poisson-inverse Gaussian regression model. It has the same framework as the models previously presented, it is, a point mass of one at zero, for handle the structural zeros, mixed with the Poisson-inverse Gaussian (PIG) distribution, introduced by Holla (1967), for handle the sampling zeros. According to Willmot (1987), the PIG distribution can be viewed as an alternative to the NB distribution and the regression model with the PIG distribution was presented by Dean et al. (1989). Hilbe (2014) presents a chapter about the Poisson-inverse Gaussian distribution and some applications and, according to him, the PIG regression is preferable when the count data have a high peak in the lower range of number and long right skewed tail, as well as for strongly Poisson overdispersed data. Hilbe (2014) also presents, in the seventh chapter, zero-inflated models and next explains the problems with zeros. The author fitted the number of visits to doctor during year, from a German health data set, several zero-inflated models, such as ZIP, ZINB and ZIPIG regression models, concluding that the ZIPIG regression models was the best-fitted model.

Other kinds of zero-inflated models have been proposed, such as the zero-inflated binomial (ZIB) regression model, introduced by Hall (2000), for upper bounded counts. The data set applied in that paper was a horticultural data set that concern establish the relationship among the number of live adults insects and some covariates. To this end, the author explains that the ZIP regression model could be adapted to the ZIB regression model, since the number of insects was bounded between zero and n. Hall (2000) obtained the parameters estimates in a similar way as Lambert (1992), using the EM algorithm. The author also proposed modification to the ZIP and the ZIB regression models, proposing an approach for these models by introducing random effects into the portion that describes the dependence of the non-zero-state mean on covariates, becoming these models useful for modeling heterogeneity and dependence in zero-inflated count data. According to the author, the assumption of independence among the responses can be violated in data sets such as the analyzed in the paper. A

table with observed values and predictions for the percentage of counts are presented for the ZIB model, where it is possible to notice the efficiency of this model, specially when a comparison is made between the ZIB regression model and the Poisson regression model and also between the ZIB regression model and the binomial logistic regression model, where both Poisson and binomial logistic regression models were under predicting the zeros count.

Essentially, we can say that count data with many zeros can be found in several fields, as mentioned by almost all of the authors, previously cited, in their works. For instance, Ridout et al. (1998) made a review wherein they cite applications in areas such as agriculture, patent applications, health care, biology, sexual behavior, road safety, use of recreational facilities, among others. Thus, in order to show the relevance of this theme and point out its importance, as well as strengthen the motivation to work with this subject, we highlight that models to handle the excess of zeros has received much attention in the literature recently and since its first appearance. To support that argument, other authors and their works can be cited besides those already mentioned.

Shankar et al. (1997) applied the ZIP regression model and the ZINB regression model to a roadway section accident data with the aim of determine which sections of the roadway was really safe, it is, those that have near zero accidents and which are not safe but have zero accidents observed during the period of observation. Böhning et al. (1999) concluded that the ZIP regression model could be considered adequate to fit a dental health care data set from a dental epidemiological study in Belo Horizonte, which evaluate programmes for reducing caries, using as response variable an important index of the dental status and as covariates age, gender, ethnicity and school.

Lee et al. (2001) made a modification to the ZIP regression model to incorporate individual exposure in the Poisson component. According to the authors, in some cases a count data is observed combined with extent of exposure and, for this reason, they generalized the ZIP regression model, with and without covariates, and applied to a manual handling injuries data set, wherein the response was the lost time injury count and the exposure was the hours worked by the orderlies from a public hospital in Western Australia.

Ridout et al. (2001) and Garay et al. (2011) applied the ZINB regression model to the apple cultivar data set reported by Ridout et al. (1998). In the paper of Ridout et al. (2001), a score test for testing the ZIP regression model against the ZINB regression model was proposed and in the paper of Garay et al. (2011), the authors report some influence diagnostics techniques, global and local ones. In a recent work, Oliveira et al. (2016) applied the ZIP and the ZINB regression models to a radiation-induced chromosome aberration data and also made a comparison among other models that handle overdispersion, such as the NB and the PIG regression models. The authors present a table resume from those models that seem to be the most appropriated to fit that kind of data.

1.1 Aims of the Thesis

The goal of this work is to deal with overdispersion and the excess of zero in count data simultaneously and, for this purpose, we will provide appropriated support for modeling these kind of data by proposing a general regression model based on a class of zero-inflated mixed Poisson distributions, exploring computational resources in the R program to obtain the model parameters estimates, as well as residual analysis and diagnostic for assess global influence.

Therefore, we have proposed a general class of zero-inflated mixed Poisson regression models to deal with overdispersion and zeros excess, which embrace some zero-inflated models, such as the ZINB regression model, and opens the opportunity to raise other models. Different of most of the works in this field, we are modeling the dispersion parameter as function of explanatory variables because the assumption that the dispersion is constant may be unrealistic in real-word cases.

On computational resources, we have proposed to obtain the model parameters estimates through the EM algorithm, which can deal with the latents variables. We have also provided the explicit expressions of the information matrix. Thus, one can obtain standard errors of the parameters estimates and, for instance, construct confidence intervals for the model parameters. With the purpose of evaluate the estimates produced by the EM algorithm, a Monte Carlo study will be presented, as well as for evaluating the estimated information matrix behavior.

In order to investigate outliers and its potential as influencer, we have made a residuals analysis based on simulated envelopes and to assess the global influence, we have used the generalized Cook's distance measure provided by Zhu et al. (2001) and, for this purpose, we provided the expressions for the zero-inflated mixed Poisson regression models of that measure, in order to check the adequacy of the assumed distribution for the response variable.

2 The Model

To make this work self-contained, in the first section we briefly present the general class of mixed Poisson (MP) distributions, which will be necessary to construct and define the class of zero-inflated mixed Poisson (ZIMP) distributions in the second one. Then, we also present in the second section the regression structure for the proposed class.

2.1 General Mixed Poisson Distributions

In order to introduce the general mixed Poisson distributions, following Barreto-Souza and Simas (2016), let Z be a continuous positive random variable belonging to the exponential family, denoting $Z \sim \text{EF}(\xi_0, \phi)$, with density function given by

$$f_Z(z) = \exp\left\{\phi\left[z\xi_0 - b\left(\xi_0\right)\right] + c\left(z;\phi\right)\right\}, \quad z > 0, \quad \phi > 0. \tag{1}$$

In addition, to define the MP class, let $Y|Z = z$ follows the Poisson distribution with μz mean, denoting $Y|Z = z \sim \text{Poisson}(\mu z)$, $\mu > 0$. In this way, a class of mixed Poisson distributions is established and we say that Y follows a general mixed Poisson distributions with μ and ϕ parameters, denoting $Y \sim \text{MP}(\mu, \phi)$. Then, its probability function is given by

$$p_Y(y;\mu,\phi) = P(Y = y) = \int_0^\infty \frac{e^{-\mu z}(\mu z)^y}{y!} f_Z(z)dz, \quad y = 0, 1, 2, \ldots. \tag{2}$$

It is possible to show that $E\left(Z\right) = b'(\xi_0)$ and $Var\left(Z\right) = \phi^{-1}b''(\xi_0)$. How proposed by the authors, here assuming the $c(z;\phi)$ function as a composition of ϕ and z functions, expressing $c(z;\phi) = d(\phi) + \phi g(z) + h(z)$. Furthermore, $b(\cdot)$ and $d(\cdot)$ are assumed being a three times differentiable functions and also ξ_0 will be determined in order to obtain $b'(\xi_0) = 1$.

Stoyanov and Lin (2011) describe the identifiability question in mixture distributions, what we briefly expose as follows. Lets (X, θ) be a two dimensional random vector defined in the $(\Omega, \Im, P)$ probability space, X taking values in the natural numbers set or a subset of natural, finite or infinite, and θ assuming values in T and $T \subset [0, \infty)$. If the conditional distribution of X given that $\theta = t$ is represented by $f(k|t) = P(X = k|\Theta = t)$, for $k = 0, 1, 2, \ldots$, and that θ has a distribution function $G(t)$, then the unconditional distribution of the random variable X, can be obtained as the mixture distribution

$$h_k = P(X = k) = \int_T f(k|t)dG(t), \quad k = 0, 1, 2, \ldots.$$

Then, the mixture distribution h_k is identifiable if, given $f(k|t)$, there is only one mixing distribution $G(\cdot)$ on T, it is, if there are two distributions on T, such that

$$h_k = P(X = k) = \int_T f(k|t)dG_1(t) = \int_T f(k|t)dG_2(t), \quad k = 0, 1, 2, \ldots,$$

$G_1 \neq G_2$, then the mixture distribution h_k is non-identifiable. According to Karlis and Xekalaki (2005), mixed Poisson distributions are identifiable, it is, in their words, *every mixed Poisson distribution corresponds to one and only one mixing distribution.*

Here, $b'(\xi_0)$ is assumed to be one not exactly because of an identifiability issue, but at first because we know that there is no loss of generality due to its choice, since mixed Poisson distributions are always identifiable and also because of a parsimonious model, otherwise the mixed Poisson distributions would depend of one more parameter. To clarify, lets begins saying that different choices for Z conduct to distinct distributions of Y. For instance, if Z follows the gamma distribution, then solving (2) one may figures out that Y follows the negative binomial distribution. Then, if Z follows the gamma

distribution with mean λ and shape parameter ϕ, the probability density function is

$$f_Z(z) = \frac{\phi^\phi}{\Gamma(\phi)\lambda^\phi} z^{\phi-1} \exp\left(-\frac{\phi}{\lambda} z\right), \quad z > 0,$$

and replacing this function in expression (2), we have

$$\begin{aligned} P(Y=y) &= \int_0^\infty \frac{e^{-\mu z}(\mu z)^y}{y!} \frac{1}{\Gamma(\phi)} \left(\frac{\phi}{\lambda}\right)^\phi z^{\phi-1} \exp\left(-\frac{\phi}{\lambda} z\right) dz \\ &= \frac{\mu^y}{y!\Gamma(\phi)} \left(\frac{\phi}{\lambda}\right)^\phi \int_0^\infty z^{y+\phi-1} \exp\left[-\left(\frac{\phi}{\lambda} + \mu\right) z\right] dz \\ &= \frac{\Gamma(y+\phi)}{y!\Gamma(\phi)} \left(\frac{\mu\lambda}{\mu\lambda+\phi}\right)^y \left(\frac{\phi}{\mu\lambda+\phi}\right)^\phi, \quad y = 0, 1, 2, \ldots, \end{aligned}$$

leading Y to follows the negative binomial distribution with $\mu\lambda$ and ϕ parameters, with mean $E(Y) = \mu\lambda$ and $Var(Y) = \mu\lambda(1 + \mu\lambda\phi^{-1})$. However, the result $\mu\lambda$ can be yield from an infinite set of combination between values of μ and λ. Thus, if we have a parameter μ^* representing the multiplication $\mu\lambda$, then we would have the same negative binomial distribution but with two parameters instead of three, what can be reached determining ξ_0 to obtain $b'(\xi_0) = 1$.

Continuing with some properties, the moment generating function of the general mixed Poisson distributions, $\varphi_Y(t) = E(e^{tY})$, can be expressed by $\varphi_Y(t) = \exp\{-\phi[b(\xi_0) - b(\xi_0 + \mu\phi^{-1}(e^t - 1))]\}$ and, with this measure, it is possible to show that the mean of Y is $E(Y) = \mu$ and its variance is $\mathrm{Var(Y)} = \mu[1 + \mu\phi^{-1}b''(\xi_0)]$. Taking a look at the variance of the general mixed Poisson distributions, it is easily noticeable that this class can handle overdispersion, since the variance is larger than the mean.

2.2 Zero-Inflated Mixed Poisson Distributions

To introduce the zero-inflated mixed Poisson distributions was necessary first to present the class of mixed Poisson distributions in the previous section and, to present the zero-inflated class, we are going to use the same structure of the last section, it is, the same nomenclature and notation, for instance, a random variable Y follows the MP distribution or, in other words, $Y \sim \text{MP}(\mu, \phi)$, as well as $Z \sim \text{EF}(\xi_0, \phi)$.

Thus, let B be distributed as a Bernoulli with probability function given by

$$P(B = b) = \tau^{1-b}(1-\tau)^b, \qquad b = 0, 1 \text{ and } 0 \leq \tau \leq 1.$$

Therefore, to set the class of zero-inflated mixed Poisson (ZIMP) distributions, we assume B and Y independent variables and define $W = BY$. Hence,

$$W = \begin{cases} 0, & \text{with } \tau \text{ probability} \\ Y, & \text{with } 1-\tau \text{ probability} \end{cases}.$$

If W belongs to the general class of zero-inflated mixed Poisson distributions, we denote $W \sim \text{ZIMP}(\mu, \phi, \tau)$ and its probability function is given by

$$p_W(w; \mu, \phi, \tau) = P(W = w) = \begin{cases} \tau + (1-\tau)\, p_Y(w; \mu, \phi), & w = 0 \\ (1-\tau) p_Y(w; \mu, \phi), & w = 1, 2, 3, \ldots \end{cases}. \tag{3}$$

The idea is to mix a class of mixed Poisson distributions with a point mass of one at zero and we say that the count of zeros of the count data is derived from two sources, some may come from the mixed Poisson distributions (or sampling zeros) and the others may come from the structural zeros, it is, that ones that do not follow or are not at the risk of the mentioned distribution, but a process ruled by a binary distribution instead.

ZIMP regression models can take some specific shape according to the distribution of its latent variable Z. For instance, if Z follows the gamma distribution with mean one and dispersion ϕ, the probability density function of Z written in terms of the exponential family is

$$f(z;\xi_0,\phi) = \exp\left\{\phi\left[z\xi_0 - (-\log(-\xi_0))\right] + \phi\log\phi - \log\Gamma(\phi) + \phi\log z - \log z\right\}, \quad z \geq 0,$$

Without loss of generality, as previous clarified, we have $\xi_0 = -1$ to obtain $b'(\xi_0) = 1$. Then, replacing the probability density function of gamma in expression (2), after some calculation, we obtain the negative binomial probability function

$$p_Y(y;\mu,\phi) = \frac{\Gamma(y+\phi)}{y!\Gamma(\phi)}\left(\frac{\mu}{\mu+\phi}\right)^y\left(\frac{\phi}{\mu+\phi}\right)^\phi, \quad y = 0,1,2,\ldots,$$

yielding $Y \sim \text{NB}(\mu,\phi)$. Thus, since W was defined as $W = BY$, then $W \sim \text{ZINB}(\mu,\phi,\tau)$, it is, $W = 0$, with τ probability, or $W = \text{NB}(\mu,\phi)$, with $1-\tau$ probability. Another example of distribution that belongs to the ZIMP distributions is when Z follows an inverse Gaussian distribution with mean one and dispersion parameter ϕ with probability density function given by

$$f(z;\xi_0,\phi) = \exp\left\{\phi\left[z\xi_0 - \left(-(-2\xi_0)^{\frac{1}{2}}\right)\right] + \frac{1}{2}\log\phi + \phi\left(\frac{-1}{2z}\right) - \frac{1}{2}\log(2\pi z^3)\right\}, \quad z \geq 0.$$

To ensure that $b'(\xi_0) = 1$, ξ_0 has been determined as $\xi_0 = -\frac{1}{2}$. Thus, after replacing the PIG distribution and solving expression (2), one may notice that Y follows the Poisson-inverse Gaussian distribution, $Y \sim \text{PIG}(\mu,\phi)$, with probability function expressed by

$$p_Y(y;\mu,\phi) = e^\phi\frac{(\mu\phi)^y}{y!}\sqrt{\frac{2}{\pi}}\left[\phi(\phi+2\mu)\right]^{-(y-1/2)/2}K_{(y-1/2)}\left(\sqrt{\phi(\phi+2\mu)}\right), \quad y = 0,1,2,\ldots,$$

Thus, since W was defined as $W = BY$, then $W \sim \text{ZIPIG}(\mu,\phi,\tau)$, it is, $W = 0$, with τ probability, or $W = \text{PIG}(\mu,\phi)$, with $1-\tau$ probability.

Some properties of ZIMP distributions can be derived by its moment generating function (mgf) $\varphi(\cdot)$. The mfg of W, denoted by $\varphi_W(\cdot)$, can be determined in function of the mgf of the general class of MP distributions. Therefore, proceeding with the calculation of Y moment generating function, denoted by $\varphi_Y(\cdot)$, after some manipulation one may notice that $\varphi_Y(t) = \exp\{-\phi\,[b(\xi_0) - b(\xi_0 + \mu\phi^{-1}(e^t - 1))]\}$. With this, the mgf of W can be derived as follows

$$\begin{aligned}
\varphi_W(t) &= E(e^{tW}) = E(e^{tBY}) = E[E(e^{tBY}|B)] \\
&= E[BE(e^{tY}) + (1 - B)] \\
&= (1 - \tau)\varphi_Y(t) + \tau \\
&= \tau + (1 - \tau)\exp\left\{-\phi\left[b(\xi_0) - b(\xi_0 + \mu\phi^{-1}(e^t - 1))\right]\right\}. \qquad (4)
\end{aligned}$$

Using the mgf of W provided in (4), it is possible to show that the mean and the variance of W are $E(W) = (1 - \tau)\mu$ and $\mathrm{Var(W)} = \mu\{1 + \mu[\phi^{-1}b''(\xi_0) + \tau]\}(1 - \tau)$, respectively. As mentioned earlier, one may figures out that MP distributions are a particular case of ZIMP distributions, when the inflation parameter τ is equal to zero. When it happens, the mgf of W is reduced to $\varphi_Y(t)$, the mgf of Y, and the mean and the variance of W are reduced to $E(W) = \mu$ and $\mathrm{Var(W)} = \mu[1 + \mu\phi^{-1}b''(\xi_0)]$ respectively, it is, the mean and the variance of Y, which follows a mixed Poisson distributions.

In order to construct the zero-inflated mixed Poisson regression models, we take into account three regression structures for the mean, the dispersion and the zero-inflation parameters. Thus, we have the following functions

$$\begin{aligned}
\log(\mu_i) &= \boldsymbol{x}_i^\top\boldsymbol{\beta}, \\
\log(\phi_i) &= \boldsymbol{v}_i^\top\boldsymbol{\alpha}, \\
\mathrm{logit}(\tau_i) &= \boldsymbol{s}_i^\top\boldsymbol{\gamma},
\end{aligned}$$

where $\boldsymbol{x}_i$, $\boldsymbol{v}_i$ and $\boldsymbol{s}_i$ are the explanatory variables vectors with $p \times 1$, $q \times 1$ and $r \times 1$ dimensions, respectively, for $i = 1, \ldots, n$, with n denoting the sample size and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)^\top$ are the parameters related to those covariates.

As previously mentioned, in many real situations the assumption that samples are independent and identically distributed is too strong and the model parameters may depend on covariates. This argument, strengthened by Cameron and Trivedi (1998), prompted us to use the regression analysis to try to understand the relationship between a count data with zeros excess and its potential explanatory variables, leading us to build regression structures for the three parameters of the zero-inflated mixed Poisson distributions. Three link functions were used, the log function, for the mean and the dispersion parameters, and the logit function, for the zero-inflation parameter. They were chosen to guaranteer the restrictions of the parameters, once the mean and the dispersion parameters are positive values and the zero-inflation parameter are restricted to the interval between zero and one. Nevertheless, other link functions that can handle the parameters restrictions could be used. In a first moment, the structure was made with three vector of covariates, however it is important to clarify that the same vector of the explanatory variables were used in the illustration chapter and we let the own model decide each one was or was not considerable.

Therefore, some zero-inflated models that can handle excess of zeros and overdispersion, that have been studied separately, are unified by the ZIMP regression models, it is, if the latent variable Z follows the gamma distribution or the inverse Gaussian distribution, than one can notices that W follows the zero-inflated negative binomial distribution and the zero-inflated Poisson-inverse Gaussian distribution, respectively, and with the regression structure we reach the zero-inflated negative binomial and the zero-inflated Poisson-inverse Gaussian regression models.

3 EM Algorithm

The EM algorithm was presented by Dempster et al. (1977) as a general approach to iterative computation of maximum likelihood estimates when the observations can be viewed as incomplete data, it is, when only a subset of the data is available. The algorithm name became from the fact that each iteration is composed of an expectation step followed by a maximization step.

In essence, the EM algorithm is used when one do not have a complete data set of observations and/or maybe the observed log-likelihood function does not have a simple constitution and, for this reason, obtain the maximum likelihood estimates can be a cumbersome process. For instance, we highlight that the log-likelihood function of the ZIPIG model, particular case of the ZIMP models, is related to the complicated modified Bessel function of the third kind. Therefore, in next paragraphs we present the steps of the algorithm.

For the zero-inflated mixed Poisson models, there are two latent variable denoted by Z, that is a distribution belonging to the continuous exponential family and mixed with the Poisson distribution to generate the mixed Poisson distribution, as mentioned in the previous chapter, and B, that is distributed as a Bernoulli(τ) distribution with τ probability at zero, related to the zero-inflation. In the general mixed Poisson models there is only Z as a latent variable.

Let $\boldsymbol{\theta}$ be the following parameters vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top)^\top$. We have worked with the complete data $(W_1, Z_1, B_1), \ldots, (W_n, Z_n, B_n)$, where $W_1, \ldots, W_n$ are the observable count data and $Z_1, \ldots, Z_n$ and $B_1, \ldots, B_n$ are the random effects. However, the Z_i's and B_i's are unobservable variables. Thus, the complete log-likelihood function is given by

$$
\begin{aligned}
l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\left\{P(W_i = w_i | Z_i = z_i, B_i = b_i) f_Z(z) P(B_i = b_i)\right\} \\
&\propto \sum_{i=1}^{n} \left\{b_i w_i \log \mu_i - b_i z_i \mu_i + b_i d(\phi_i) + \phi_i \left[b_i z_i \xi_0 - b_i b(\xi_0) + b_i g(z_i)\right]\right. \\
&\qquad \left. + b_i \text{logit}(1-\tau_i) + \log(\tau_i)\right\}. \qquad (5)
\end{aligned}
$$

Two steps are required to carry out the EM algorithm, the E-step (expectation step) and the M-step (maximization step). The goal of the first one is to compute the complete log-likelihood function conditional expectation, which defines the Q function. In the second one, the aim is maximize the Q function. If we denote $\boldsymbol{\theta}^{(0)}$ as the initial $\boldsymbol{\theta}$ estimate, the Q function is updated and an estimate $\boldsymbol{\theta}^{(1)}$ is obtained as the argument which maximizes Q and this is the first-step estimate of $\boldsymbol{\theta}$. Then, this procedure is performed as much as needed to some criterion convergence be reached. The algorithm is described in details in what follows.

Expectation Step

Let $\boldsymbol{\theta}^{(r)}$ be the estimate of $\boldsymbol{\theta}$ on the rth step. First of all, it is necessary to compute the Q function as follows

$$
\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) &= E(l_c(\boldsymbol{\theta}) | \mathbf{W}; \boldsymbol{\theta}^{(r)}) \\
&\propto \sum_{i=1}^{n} \Big\{ w_i \log \mu_i \delta_i^{(r)} - \mu_i \lambda_i^{(r)} + d(\phi_i)\delta_i^{(r)} + \phi_i [\xi_0 \lambda_i^{(r)} - \delta_i^{(r)} b(\xi_0) \\
&\qquad + \kappa_i^{(r)}] + \delta_i^{(r)} \text{logit}(1-\tau_i) + \log(\tau_i) \Big\}, \qquad (6)
\end{aligned}
$$

where

$$\begin{aligned}
\delta_i^{(r)} &= E(B_i|\mathbf{W};\boldsymbol{\theta}^{(r)}),\\
\lambda_i^{(r)} &= E(B_i Z_i|\mathbf{W};\boldsymbol{\theta}^{(r)}),\\
\kappa_i^{(r)} &= E(B_i g(Z_i)|\mathbf{W};\boldsymbol{\theta}^{(r)}).
\end{aligned}$$

The results of the previous conditional expectations are given in the following proposition.

Proposition 1 *Let $W \sim ZIMP(\mu,\phi,\tau)$, with $Z \sim EF(\xi_0,\phi)$ and $B \sim Bernoulli(\tau)$, the previous latent variables defined, and $Y \sim MP(\mu,\phi)$. Thus,*

$$E(B|W) = (1-\tau)\frac{p_Y(w;\mu,\phi)}{p_W(w;\mu,\phi,\tau)}, \tag{7}$$

$$\begin{aligned}
E(BZ|W) &= (1-\tau)\frac{p_Y(w;\mu,\phi)}{p_W(w;\mu,\phi,\tau)}(w+1)\frac{p_Y(w+1,\mu,\phi)}{\mu p_Y(w,\mu,\phi)}\\
&= \frac{(1-\tau)(w+1)}{\mu}\frac{p_Y(w+1;\mu,\phi)}{p_W(w;\mu,\phi,\tau)},
\end{aligned} \tag{8}$$

$$E(Bg(Z)|W) = (1-\tau)\frac{p_Y(w;\mu,\phi)}{p_W(w;\mu,\phi,\tau)}\left(\frac{dp_Y(w;\mu_t^*,\phi+t)/dt|_{t=0}}{p_Y(w;\mu,\phi)} - d'(\phi) - \xi_0 + b(\xi_0)\right), \tag{9}$$

where $\mu^* = b'\left(\dfrac{\phi\xi_0}{\phi+t}\right)$. When we are at the case that there is no zero-inflation, it is, when the zero-inflation parameter τ is equal to zero, then W is reduced to Y. If its happen, then the term $(1-\tau)\dfrac{p_Y(w;\mu,\phi)}{p_W(w;\mu,\phi,\tau)}$ goes to one and disappear, while and the equations (8) and (9) are reduced to $E(Z|Y)$ and $E(g(Z)|Y)$, respectively, the conditional expectations of the complete log-likelihood function in the class of mixed Poisson regression models. Because of that, we can easily notice that the MP regression models are particular cases of the zero-inflated mixed Poisson regression models proposed in this work.

Remark: The proof of Proposition 1 is presented in the Appendix.

Maximization Step

In order to improve the algorithm to obtain the argument that maximizes the Q function, it is necessary to obtain the score function associated to the Q function, given by

$$\frac{\partial Q}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \delta_i^{(r)} w_i - \mu_i \lambda_i^{(r)} \right\} x_{ij}, \qquad j = 1, \ldots, p; \tag{10}$$

$$\frac{\partial Q}{\partial \alpha_l} = \sum_{i=1}^{n} \phi_i \left\{ \xi_0 \lambda_i^{(r)} + \delta_i^{(r)} [d'(\phi_i) - b(\xi_0)] + \kappa_i^{(r)} \right\} y_{il}, \qquad l = 1, \ldots, q; \tag{11}$$

$$\frac{\partial Q}{\partial \gamma_m} = \sum_{i=1}^{n} \left\{ 1 - \tau_i - \delta_i^{(r)} \right\} s_{im}, \qquad m = 1, \ldots, r. \tag{12}$$

Therefore, the estimates can be obtained updating $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ with $\delta_i^{(r)}$, $\lambda_i^{(r)}$ and $\kappa_i^{(r)}$ through the estimate $\boldsymbol{\theta}^{(r)}$ in the rth step. Then, $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ needs to be maximized under $\boldsymbol{\theta}$ and it can be done through a numerical optimization algorithm (in this work the method Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm has been applied) and improved by using the score function. This routine will be repeated until some convergence criterion be reached, for instance $\| \boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)} \| < \epsilon$, $\| \boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)} \| / \| \boldsymbol{\theta}^{(r)} \| < \epsilon$ or $\| Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r+1)}) - Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) \| < \epsilon$. In this work, a combination between the first one and the last one has been used taking into account $\epsilon = 10^{-4}$.

3.1 Information Matrix

According to Louis (1982), the observed information matrix when one uses the EM algorithm, here denoted as $I(\boldsymbol{\theta})$, is given by

$$I(\boldsymbol{\theta}) = E\left(-\frac{\partial l_c(\boldsymbol{\theta})^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} | \boldsymbol{W} \right) - E\left(\frac{\partial l_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial l_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\top | \boldsymbol{W} \right). \tag{13}$$

The elements of the observed information matrix (13), presented below, have been used to obtain the standard errors of the parameters estimates, as it is possible to see in Empirical Illustration chapter. In the Simulation Study chapter a Monte Carlo study is presented in order to show the finite sample behavior of the estimated information matrix $I(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ obtained thorough the EM algorithm. Thus, the elements of the observed information matrix (13) are obtained by

$$E\left(-\frac{\partial l_c^2}{\partial\beta_j\partial\beta_l}|\boldsymbol{W}\right)=\sum_{i=1}^{n}\lambda_i\mu_i x_{ij}x_{il},\quad \text{for } j,l=1,\ldots,p;$$

$$E\left(-\frac{\partial l_c^2}{\partial\alpha_j\partial\alpha_l}|\boldsymbol{W}\right)=\sum_{i=1}^{n}\phi_i\left\{\delta_i\left[b(\xi_0)-d'(\phi_i)-d''(\phi_i)\phi_i\right]-\lambda_i\xi_0-\kappa_i\right\}v_{ij}v_{il},$$

for $j,l=1,\ldots,q$;

$$E\left(-\frac{\partial l_c^2}{\partial\gamma_j\partial\gamma_l}|\boldsymbol{W}\right)=\sum_{i=1}^{n}\tau_i(1-\tau_i)s_{ij}s_{il},\quad \text{for } j,l=1,\ldots,r;$$

$$E\left(-\frac{\partial l_c^2}{\partial\beta_j\partial\alpha_l}|\boldsymbol{W}\right)=0,\quad \text{for } j=1,\ldots,p \text{ and } l=1,\ldots,q;$$

$$E\left(-\frac{\partial l_c^2}{\partial\beta_j\partial\gamma_l}|\boldsymbol{W}\right)=0,\quad \text{for } j=1,\ldots,p \text{ and } l=1,\ldots,r;$$

$$E\left(-\frac{\partial l_c^2}{\partial\gamma_j\partial\alpha_l}|\boldsymbol{W}\right)=0,\quad \text{for } j=1,\ldots,r \text{ and } l=1,\ldots,q;$$

$$
\begin{aligned}
E\left(\frac{\partial l_c}{\partial \beta_j}\frac{\partial l_c}{\partial \beta_l}|\boldsymbol{W}\right) &= \sum_{i=1}^{n}\{\psi_i w_i^2 - 2\zeta_i\mu_i w_i + \eta_i\mu_i^2\}x_{ij}x_{il} \\
&+ \sum_{i\neq k}(\delta_i w_i - \lambda_i\mu_i)(\delta_k w_k - \lambda_k\mu_k)x_{ij}x_{kl},
\end{aligned}
$$

for $j, l = 1, \ldots, p$;

$$
\begin{aligned}
E\left(\frac{\partial l_c}{\partial \alpha_j}\frac{\partial l_c}{\partial \alpha_l}|\boldsymbol{W}\right) &= \sum_{i=1}^{n}\phi_i^2\{\eta_i\xi_0^2 + 2\rho_i\xi_0 - 2[\zeta_i\xi_0 + \upsilon_i][b(\xi_0) - d'(\phi_i)] \\
&+ \psi_i[b(\xi_0) - d'(\phi_i)]^2 + \nu_i\}v_{ij}v_{il} \\
&+ \sum_{i\neq k}\phi_i\phi_k\left\{\lambda_i\xi_0 + \kappa_i - \delta_i\left[b(\xi_0) - d'(\phi_i)\right]\right\}\left\{\lambda_k\xi_0 + \kappa_k\right. \\
&- \left.\delta_k\left[b(\xi_0) - d'(\phi_k)\right]\right\}v_{ij}v_{kl},
\end{aligned}
$$

for $j, l = 1, \ldots, q$;

$$
\begin{aligned}
E\left(\frac{\partial l_c}{\partial \gamma_j}\frac{\partial l_c}{\partial \gamma_l}|\boldsymbol{W}\right) &= \sum_{i=1}^{n}\{(1-\tau_i)^2 - 2(1-\tau_i)\delta_i + \psi_i\}s_{ij}s_{il} \\
&+ \sum_{i\neq k}(1 - \tau_i - \delta_i)(1 - \tau_k - \delta_k)s_{ij}s_{kl},
\end{aligned}
$$

for $j, l = 1, \ldots, r$;

$$
\begin{aligned}
E\left(\frac{\partial l_c}{\partial \beta_j}\frac{\partial l_c}{\partial \alpha_l}|\boldsymbol{W}\right) &= \sum_{i=1}^{n}\phi_i\{\zeta_i[w_i\xi_0 - \mu_i[d'(\phi_i) - b(\xi_0)]] \\
&+ [\upsilon_i + \psi_i[d'(\phi_i) - b(\xi_0)]]w_i - [\eta_i\xi_0 + \rho_i]\mu_i\}x_{ij}v_{il} \\
&+ \sum_{i\neq k}\phi_k\left\{[\delta_i w_i - \lambda_i\mu_i][\lambda_k\xi_0 + \kappa_k + \delta_k[d'(\phi_k) - b(\xi_0)]]\right\}x_{ij}v_{kl},
\end{aligned}
$$

for $j = 1, \ldots, p$ and $l = 1, \ldots, q$;

$$
\begin{aligned}
E\left(\frac{\partial l_c}{\partial \beta_j}\frac{\partial l_c}{\partial \gamma_l}|\boldsymbol{W}\right) &= \sum_{i=1}^{n}\{[(1-\tau_i)\delta_i - \psi_i]w_i - [(1-\tau_i)\lambda_i - \zeta_i]\mu_i\}x_{ij}s_{il} \\
&+ \sum_{i\neq k}(\delta_i w_i - \lambda_i\mu_i)(1-\tau_k-\delta_k)x_{ij}s_{kl},
\end{aligned}
$$

for $j = 1, \ldots, p$ and $l = 1, \ldots, r$;

$$
\begin{aligned}
E\left(\frac{\partial l_c}{\partial \gamma_j}\frac{\partial l_c}{\partial \alpha_l}|\boldsymbol{W}\right) &= \sum_{i=1}^{n}\phi_i\{(1-\tau_i)[\lambda_i\xi_0 + \kappa_i + \delta_i[d'(\phi_i) - b(\xi_0)]] \\
&- \zeta_i\xi_0 - \upsilon_i - \psi_i[d'(\phi_i) - b(\xi_0)]\}s_{ij}v_{il} \\
&+ \sum_{i\neq k}\phi_k(1-\tau_i-\delta_i)\left\{\lambda_k\xi_0 + \kappa_k + \delta_k[d'(\phi_k) - b(\xi_0)]\right]\}s_{ij}v_{kl},
\end{aligned}
$$

for $j = 1, \ldots, r$ and $l = 1, \ldots, q$.

Here, δ_i, λ_i and κ_i are defined as in **Propositon 1** and ψ_i, ζ_i, η_i, υ_i, ν_i and ρ_i will be defined as

$$
\begin{aligned}
\psi_i &= E(B_i^2|\mathbf{W};\boldsymbol{\theta}^{(r)}), \\
\zeta_i &= E(B_i^2 Z_i|\mathbf{W};\boldsymbol{\theta}^{(r)}), \\
\eta_i &= E(B_i^2 Z_i^2|\mathbf{W};\boldsymbol{\theta}^{(r)}), \\
\upsilon_i &= E(B_i^2 g(Z_i)|\mathbf{W};\boldsymbol{\theta}^{(r)}), \\
\nu_i &= E(B_i^2 g^2(Z_i)|\mathbf{W};\boldsymbol{\theta}^{(r)}), \\
\rho_i &= E(B_i^2 Z_i g(Z_i)|\mathbf{W};\boldsymbol{\theta}^{(r)}).
\end{aligned}
$$

The explicit expressions for the conditional expectations above will be given in Appendix.

3.2 Residuals

The cycle of model specification includes estimation, testing and evaluation to analyze a count data. For the last step, one might perform, for instance, residuals analysis and use goodness-of-fit measures.

According to Cameron and Trivedi (1998), a residuals analysis can be used for several objectives such as to detect model misspecification, outliers, poor fit observations, influential observations or those ones that produce a big impact on the fitted model. In other words, residuals analysis measures the departure of fitted values from actual values of the dependent variable and a visual analysis may potentially indicates the nature of misspecification and the magnitude of its effect.

The raw residual is defined as the difference between the actual and the fitted value and, for the classical linear regression model the raw residual is, asymptotically, symmetrically distributed around zero with constant variance. However, it is not true for count data, it is, the residuals do not necessarily have zero mean, constant variance or symmetric distribution. Thus, the Pearson residual is a correction for the heteroscedasticity and it is defined as

$$r_i = \frac{w_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}^2}}, \tag{14}$$

where

$$\hat{\mu}_i = \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{s}_i^\top \hat{\boldsymbol{\gamma}})},$$

$$\hat{\sigma}^2 = \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \left\{ 1 + \left[\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) - \frac{\exp(\boldsymbol{s}_i^\top \hat{\boldsymbol{\gamma}})}{1 + \exp(\boldsymbol{s}_i^\top \hat{\boldsymbol{\gamma}})} \right] \left[\frac{b''(\xi_0)}{\exp(\boldsymbol{v}_i^\top \hat{\boldsymbol{\alpha}})} + \frac{\exp(\boldsymbol{s}_i^\top \hat{\boldsymbol{\gamma}})}{1 + \exp(\boldsymbol{s}_i^\top \hat{\boldsymbol{\gamma}})} \right] \right\}$$

and $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ are the maximum likelihood estimates (MLE's) of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively.

One may expect that the residuals be concentrated around zero, but they do not follow the normal distribution. Therefore, a way to take it into account and check the model adequacy is to use simulated envelopes, which takes account of the overdispersion, as pointed by Hinde and Demétrio (1998). Then, the algorithm to build the simulated envelope is presented at Algorithm 1.

Algorithm 1 - Simulated envelope for residuals

1. For each $i = 1, \ldots, n$, compute $\hat{\mu}_i$, $\hat{\phi}_i$ and $\hat{\tau}_i$.
2. Generate n observations $\tilde{W}_i$, where $\tilde{W}_i \sim \text{ZIMP}(\hat{\mu}_i, \hat{\phi}_i, \hat{\tau}_i)$.
3. Obtain the regression coefficients $\tilde{\theta} = (\tilde{\beta}^\top, \tilde{\alpha}^\top, \tilde{\gamma}^\top)^\top$ from the regression of $\tilde{W}$ on the covariates.
4. Compute Pearson residual using $\tilde{W}_i$ and expression (14) and denote the resulting residual by $\tilde{R}_i$.
5. Repeat the previous steps m times, thus obtaining m residuals $\tilde{R}_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.
6. For each j, sort the n residuals in non-decreasing order, obtaining $\tilde{R}_{(i)j}$.
7. For i, obtain the percentiles 2.5% and the 97.5% of the ordered residuals $\tilde{R}_{(i)j}$ over j: $\tilde{R}_i^{2.5\%}$ and $\tilde{R}_i^{97.5\%}$ respectively.
8. The lower and the upper bounds for each residual R_i of the original regression are given by $\tilde{R}_i^{2.5\%}$ and $\tilde{R}_i^{97.5\%}$, respectively.

3.3 Diagnostics

Outliers are defined as the value of some point that is very distinct from the value predicted by the regression model. In other words, an outlier is the observation that has large residual. However, an outlier is not necessary an influential point, it is, an observation that pursue a large influence on the fit of the model. For this reason, the residuals analysis might not evaluate the impact that an observation may causes in the

estimation and the inference of the model parameters.

For such purpose, in this work we focus on the global influence to measure the impact that some observation may causes in the model fit. One method to find out influential points is comparing the fit of the model with and without each observation. Thus, we are going to use the generalized Cook's distance, proposed by Zhu et al. (2001), that measures, in a general way, the influence of each observation of the regression coefficients, it is, they generalized the statistic prosed by Cook (1977) as a measure of the extent of change in model estimates when a particular observation is omitted. In other words, the idea is to compare the difference between the maximum likelihood estimates with and without an observation and observe how far they are each other. If the deletion of that observation seriously impact the estimates, then that particular point requires more investigation.

Zhu et al. (2001) obtained the generalized Cook's distance (GCD) measure on bases of the complete log-likelihood function conditional expectation for models with incomplete data

$$GCD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[i]} - \hat{\boldsymbol{\theta}})^\top \{-\ddot{Q}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})\}(\hat{\boldsymbol{\theta}}_{[i]} - \hat{\boldsymbol{\theta}}),$$

where $\ddot{Q}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}) = \dfrac{\partial^2 Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and a measure with the subscript $[i]$ indicates a quantity that was computed without the ith observation deleted. In order to avoid compute all $\hat{\boldsymbol{\theta}}_{[i]}$ cases, what can be a cumbersome task, the authors provided the following one step approximation $\hat{\boldsymbol{\theta}}^1_{[i]}$ of $\hat{\boldsymbol{\theta}}_{[i]}$

$$\hat{\boldsymbol{\theta}}^1_{[i]} = \hat{\boldsymbol{\theta}} + \{-\ddot{Q}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})\}^{-1} \dot{Q}_{[i]}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}),$$

where $\dot{Q}_{[i]}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}) = \dfrac{\partial Q_{[i]}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. In what follows, we present the expressions of the

ZIMP models for that approximation

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^1_{[i]} &= \hat{\boldsymbol{\beta}} + \{(\boldsymbol{X}^\top \boldsymbol{G_1} \boldsymbol{X})^{-1} \boldsymbol{a}_i \boldsymbol{x}_i\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \\
\hat{\boldsymbol{\alpha}}^1_{[i]} &= \hat{\boldsymbol{\alpha}} + \{(\boldsymbol{V}^\top \boldsymbol{G_2} \boldsymbol{V})^{-1} \boldsymbol{b}_i \boldsymbol{v}_i\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \\
\hat{\boldsymbol{\gamma}}^1_{[i]} &= \hat{\boldsymbol{\gamma}} + \{(\boldsymbol{S}^\top \boldsymbol{G_3} \boldsymbol{S})^{-1} \boldsymbol{c}_i \boldsymbol{s}_i\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.
\end{aligned}$$

where $\boldsymbol{a}_i = \delta_i w_i - \lambda_i \mu_i$, $\boldsymbol{b}_i = \phi_i\{\xi_0 \lambda_i + \kappa_i + \delta_i[d'(\phi_i) - b(\xi_0)]\}$, $\boldsymbol{c}_i = 1 - \tau_i - \delta_i$, $\boldsymbol{G_1} = \text{diag}(\mu_i \lambda_i)$, $\boldsymbol{G_2} = \text{diag}(\delta_i[b(\xi_0) - d'(\phi_i) - d''(\phi_i)] - \xi_0 \lambda_i - \kappa_i)$ and $\boldsymbol{G_3} = \text{diag}(\tau_i(1-\tau_i))$, for $i = 1, \ldots, n$. Therefore, those approximations can be used to obtain one step approach of the GCD as

$$\begin{aligned}
GCD^1_i(\boldsymbol{\theta}) &= \dot{Q}_{[i]}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})^\top \{-\ddot{Q}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})\}^{-1} \dot{Q}_{[i]}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}) \\
&= \left\{\boldsymbol{a}_i^2 \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{G_1} \boldsymbol{X})^{-1} \boldsymbol{x}_i\right\} \\
&+ \left\{\boldsymbol{b}_i^2 \boldsymbol{v}_i^\top (\boldsymbol{V}^\top \boldsymbol{G_2} \boldsymbol{V})^{-1} \boldsymbol{v}_i\right\} \\
&+ \left\{\boldsymbol{c}_i^2 \boldsymbol{s}_i^\top (\boldsymbol{S}^\top \boldsymbol{G_3} \boldsymbol{S})^{-1} \boldsymbol{s}_i\right\}. \qquad (15)
\end{aligned}$$

Furthermore, expression (15) can be divided in three components, the $\boldsymbol{\beta}$ component, the $\boldsymbol{\alpha}$ component and the $\boldsymbol{\tau}$ one, as we respectively reported below

$$\begin{aligned}
GCD^1_i(\boldsymbol{\beta}) &= \left\{\boldsymbol{a}_i^2 \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{G_1} \boldsymbol{X})^{-1} \boldsymbol{x}_i\right\}, \\
GCD^1_i(\boldsymbol{\alpha}) &= \left\{\boldsymbol{b}_i^2 \boldsymbol{v}_i^\top (\boldsymbol{V}^\top \boldsymbol{G_2} \boldsymbol{V})^{-1} \boldsymbol{v}_i\right\}, \\
GCD^1_i(\boldsymbol{\gamma}) &= \left\{\boldsymbol{c}_i^2 \boldsymbol{s}_i^\top (\boldsymbol{S}^\top \boldsymbol{G_3} \boldsymbol{S})^{-1} \boldsymbol{s}_i\right\}.
\end{aligned}$$

In the next chapter, a Monte Carlo study will be exhibited in order to check the finite sample maximum likelihood estimates performance, produced by the EM algorithm previously introduced and, in the Empirical Illustration chapter the measures of residuals analysis and diagnostics will be applied.

4 Simulation Study

A simulation study was performed in order to check the results obtained through the EM algorithm for finite samples. Three simulation scenarios were performed, considering a low zero-inflation rate, an average zero-inflation rate and a high zero-inflation rate, using the R program (R Core Team 2016) and taking into account 4500 runs of Monte Carlo and with the following regression structure for the mean, the dispersion and the zero-inflation parameters

- $\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{4i}$,
- $\log(\phi_i) = \alpha_0 + \alpha_1 x_{2i} + \alpha_2 x_{4i}$,
- $\text{logit}(\tau_i) = \gamma_0 + \gamma_1 x_{3i} + \gamma_2 x_{4i}$,

for $i = 1, \ldots, n$, with x_{1i}, x_{2i} and x_{3i} generated independently from standard uniform distributions and x_{4i} generated independently from the Poisson distribution with a 0.5 mean rate, all variables fixed throughout the simulation.

For all scenarios were performed samples of sizes $n = 50,\ 100,\ 200,\ 300$. In the first one, we set $(\beta_0,\ \beta_1,\ \beta_2) = (1.0,\ 1.0,\ 1.5)$, $(\alpha_0,\ \alpha_1,\ \alpha_2) = (0.5,\ 1.5,\ 1.0)$ and $(\gamma_0,\ \gamma_1,\ \gamma_2) = (-2.0,\ 1.0,\ -1.5)$, providing ranges equal to $(2.72,\ 13,359.73)$, $(1.65,\ 1,096.63)$ and $(2.0\times10^{-4},\ 0.12)$ for μ, ϕ and τ, respectively, and an average zero-inflation rate of 10%. In the second and third scenarios we set almost the same values, however $(\gamma_0,\ \gamma_1,\ \gamma_2)$ has been modified to $(0.5,\ -1.5,\ -2.0)$ and $(1.0,\ -0.5,\ -1.5)$ with the purpose of get a moderate and a higher zero-inflation rate, approximately 30% and 50%, respectively. With this, the range of τ is $(1.67\times10^{-5},\ 0.62)$ and $(9.0\times10^{-4},\ 0.73)$ for the moderate and the high zero-inflation scenarios, respectively.

In each scenario, we set the zero-inflated negative binomial regression model and the zero-inflated Poisson-inverse Gaussian regression model, two particular cases from the general class of zero-inflated mixed Poisson regression models. We generate the count data from the negative binomial distribution and the Poisson-inverse Gaussian distribution, both with mean μ_i and dispersion parameter ϕ_i and then they were inflated through a Bernoulli with τ_i probability at zero. Therefore, yielding $w_i \sim \text{ZINB}(\mu_i, \phi_i, \tau_i)$ and $w_i \sim \text{ZIPIG}(\mu_i, \phi_i, \tau_i)$ for each case.

Hereinafter, the results of the three scenarios for the ZINB and the ZIPIG regression models will be presented in tables 1, 2 and 3 for the ZINB case and table 5 for ZIPIG case, wherein it is possible to check the EM algorithm estimates performance through the mean and the root of the mean square error(RMSE) of the estimated parameters. We remark that a table with the EM algorithm time to performs the simulation was reported in the appendix.

We have also made a comparison between the estimates of the proposed ZIMP regression models and the generalized additive models for location, scale and shape (GAMLSS). According to Stasinopoulos and Rigby (2007), GAMLSS is a structure for fitting regression models wherein the premise that the distribution of the response variable belongs to the exponential family is relaxed, being replaced by highly skew and kurtotic continuous and discrete distributions. GAMLSS allows modeling the mean and other parameters, being used to model a response variable that does not follows a distribution belonging to the exponential family and also to deal with heterogeneity. Among the distributions that belong to GAMLSS, we can cite the ZINB and the ZIPIG distributions. The difference between GAMLSS and the proposed ZIMP regression models is that here an EM algorithm has been proposed to obtain the maximum likelihood estimates of the parameters, while in GAMLSS they are direct obtained by maximizing the likelihood function. The results presented hereafter strengthen the proposed EM algorithm.

Table 1: Mean and root of the mean square error, in parentheses, of the parameters estimates for the ZINB model - 10% zero-inflation scenario

Sample size	$n = 50$		$n = 100$	
Θ	ZINB	GAMLSS	ZINB	GAMLSS
β_0	0.987 (0.388)	0.990 (0.389)	0.988 (0.275)	0.989 (0.275)
β_1	0.995 (0.524)	0.993 (0.526)	1.004 (0.308)	1.005 (0.309)
β_2	1.512 (0.453)	1.510 (0.455)	1.506 (0.312)	1.505 (0.313)
α_0	0.737 (1.872)	0.856 (2.859)	0.571 (1.160)	0.579 (1.204)
α_1	1.563 (2.478)	1.749 (3.289)	1.584 (1.449)	1.626 (1.532)
α_2	1.013 (2.477)	0.829 (3.526)	1.077 (1.571)	1.055 (1.634)
γ_0	-3.186 (6.729)	-1×10^{12} (7×10^{13})	-2.195 (1.665)	-2.189 (1.630)
γ_1	1.469 (2.726)	-4×10^{12} (2×10^{14})	1.089 (0.775)	-4×10^{11} (4×10^{13})
γ_2	-1.603 (10.442)	-2×10^{12} (1×10^{14})	-1.623 (2.512)	-1.623 (2.496)
Sample size	$n = 200$		$n = 300$	
Θ	ZINB	GAMLSS	ZINB	GAMLSS
β_0	0.995 (0.172)	0.996 (0.172)	0.996 (0.135)	0.997 (0.135)
β_1	1.001 (0.199)	1.002 (0.200)	1.001 (0.155)	1.002 (0.155)
β_2	1.505 (0.198)	1.504 (0.198)	1.502 (0.157)	1.501 (0.156)
α_0	0.510 (0.742)	0.519 (0.759)	0.516 (0.595)	0.527 (0.608)
α_1	1.578 (0.996)	1.605 (1.022)	1.532 (0.798)	1.557 (0.822)
α_2	1.044 (0.997)	1.020 (1.020)	1.034 (0.777)	1.007 (0.793)
γ_0	-2.098 (0.813)	-196.881 (2×10^4)	-2.068 (0.613)	-2.056 (0.607)
γ_1	1.049 (0.457)	98.445 (9×10^3)	1.034 (0.343)	1.032 (0.343)
γ_2	-1.560 (1.362)	-1.576 (1.362)	-1.516 (1.042)	-1.534 (1.040)

Taking a look at table 1, in general the estimates obtained through the EM algorithm performs well about bias and RMSE criteria, even for small samples size, except for γ's parameters on 50 samples size, which has -3.186 as estimate for the right -2 value of γ_0 and 10.442 as RMSE value for γ_2, a large value when comparing with the others. Comparing the EM estimates of the proposed model with the estimates produced by the GAMLSS, one can notice that GAMLSS performs well as EM only for the samples of size 300, but when one take a look at the smaller samples size it is possible to observe the γ's parameters extremely biased and with high RMSE values, as well as a not so good performance of α's parameters for samples of size 50.

Therefore, the EM algorithm estimates performed low bias, mainly for moderate or large samples size. Briefly describing the EM algorithm estimates performance for RMSE criterion, it performs well for almost all parameters, except for γ's on 50 samples size, however RMSE fast decreases when samples size increases.

Analyzing table 2, we see a similar pattern about the bias and the RMSE criteria, just as the results presented in table 1. Furthermore, in the second scenario, which has a moderate zero-inflation rate, around 30%, we observe a better performance than in the first one, since the bias and the RMSE are lower, even for γ's parameters on 50 samples size.

By looking the GAMLSS results, just as for the scenario of the 10% zero-inflation rate, one can see bad performance for γ_1 parameter in any of the samples size, even for samples of size 300. This parameter is related to the variable generated according to the Poisson distribution with a 0.5 mean rate, but it did not affect the estimates of the proposed EM algorithm, that performs well even for small samples size.

Table 3 holds the scenario with a high zero-inflation rate and allows one to notice that the EM algorithm produced good estimates for all samples size, since we can realize low bias and RMSE, whereas the GAMLSS did not performs well for γ_1 parameter on samples of size 50 and 100, producing estimates completely biased and with high RMSE values. In summary, the GAMLSS performs well for the high zero-inflation scenario for 200 samples size or higher.

Table 2: Mean and root of the mean square error, in parentheses, of the parameters estimates for the ZINB model - 30% zero-inflation scenario

Sample size	$n = 50$		$n = 100$	
Θ	ZINB	GAMLSS	ZINB	GAMLSS
β_0	0.974 (0.498)	0.970 (0.506)	0.983 (0.298)	0.982 (0.299)
β_1	1.004 (0.449)	1.006 (0.453)	1.006 (0.289)	1.008 (0.290)
β_2	1.523 (0.518)	1.527 (0.528)	1.514 (0.340)	1.513 (0.341)
α_0	0.706 (1.903)	0.751 (3.399)	0.597 (1.328)	0.598 (1.394)
α_1	1.540 (2.558)	1.678 (3.817)	1.579 (1.610)	1.628 (1.686)
α_2	1.067 (2.445)	1.154 (3.848)	1.052 (1.761)	1.036 (1.857)
γ_0	0.448 (1.259)	0.397 (1.435)	0.523 (0.870)	0.519 (0.868)
γ_1	-2.444 (4.200)	$-3\times10^{10}(1\times10^{15})$	-1.787 (1.692)	$-2\times10^{13}(2\times10^{14})$
γ_2	-2.118 (2.102)	-2.060 (2.239)	-2.069 (1.468)	-2.063 (1.465)
Sample size	$n = 200$		$n = 300$	
Θ	ZINB	GAMLSS	ZINB	GAMLSS
β_0	0.995 (0.203)	0.995 (0.203)	0.996 (0.164)	0.997 (0.164)
β_1	0.997 (0.217)	0.997 (0.217)	1.002 (0.172)	1.002 (0.172)
β_2	1.505 (0.224)	1.505 (0.225)	1.502 (0.184)	1.500 (0.184)
α_0	0.520 (0.877)	0.530 (0.909)	0.511 (0.681)	0.527 (0.692)
α_1	1.574 (1.077)	1.597 (1.110)	1.531 (0.836)	1.550 (0.857)
α_2	1.045 (1.070)	1.025 (1.107)	1.043 (0.892)	1.013 (0.905)
γ_0	0.497 (0.527)	0.498 (0.525)	0.495 (0.437)	0.499 (0.436)
γ_1	-1.587 (0.715)	$-2\times10^{12}(5\times10^{13})$	-1.553 (0.483)	$-5\times10^{10}(4\times10^{12})$
γ_2	-2.019 (0.897)	-2.020 (0.895)	-2.008 (0.768)	-2.014 (0.768)

Before presenting the results of the ZIPIG case, we may first check how some parameters estimates of the 4500 samples are distributed, making a comparison between the estimates of the ZIMP models and the GAMLSS.

Table 3: Mean and root of the mean square error, in parentheses, of the parameters estimates for the ZINB model - 50% zero-inflation scenario

Sample size	$n = 50$		$n = 100$	
Θ	ZINB	GAMLSS	ZINB	GAMLSS
β_0	0.957 (0.638)	0.949 (0.659)	0.980 (0.369)	0.978 (0.373)
β_1	1.010 (0.592)	1.014 (0.603)	1.008 (0.407)	1.010 (0.410)
β_2	1.541 (0.749)	1.549 (0.776)	1.512 (0.447)	1.514 (0.451)
α_0	0.744 (2.646)	0.779 (4.787)	0.642 (1.653)	0.634 (2.110)
α_1	1.116 (2.939)	1.279 (5.332)	1.499 (2.035)	1.612 (2.653)
α_2	1.303 (3.260)	1.393 (5.477)	1.123 (2.152)	1.154 (3.102)
γ_0	0.892 (1.337)	0.835 (1.494)	1.010 (0.693)	1.003 (0.702)
γ_1	-0.607 (0.955)	-2×10^{12} (6×10^{13})	-0.536 (0.528)	-2×10^{11} (1×10^{13})
γ_2	-1.407 (2.016)	-1.340 (2.193)	-1.547 (1.255)	-1.539 (1.269)
Sample size	$n = 200$		$n = 300$	
Θ	ZINB	GAMLSS	ZINB	GAMLSS
β_0	0.991 (0.241)	0.992 (0.241)	0.995 (0.205)	0.995 (0.205)
β_1	1.000 (0.268)	1.001 (0.268)	0.997 (0.225)	0.998 (0.225)
β_2	1.507 (0.292)	1.505 (0.292)	1.505 (0.233)	1.504 (0.233)
α_0	0.555 (1.132)	0.570 (1.172)	0.535 (0.913)	0.550 (0.932)
α_1	1.548 (1.333)	1.581 (1.395)	1.563 (1.110)	1.581 (1.138)
α_2	1.069 (1.479)	1.039 (1.519)	1.026 (1.121)	1.000 (1.139)
γ_0	1.003 (0.462)	1.004 (0.462)	1.006 (0.392)	1.007 (0.392)
γ_1	-0.511 (0.349)	-0.511 (0.349)	-0.514 (0.267)	-0.514 (0.266)
γ_2	-1.508 (0.772)	-1.510 (0.772)	-1.505 (0.622)	-1.508 (0.623)

In figure 1, through the boxplots it is possible to verify that GAMLSS produced several outliers for the α's parameters, what does not happen with the EM algorithm estimates. The analysis of γ_1, for $n = 50$, shows the same behavior of α's, helping us to understand the reason of such bad estimates presented in table 1. However, it is important to point that the bad estimates of γ_0, for $n = 50$ and $n = 200$, occur because of some few extreme outliers (one or two), which leads us to believe in better estimates after removing the corresponding samples, but it is not the case for samples of size 100, once its has some outliers but not with extreme high values. Moreover, the EM algorithm estimates performs similar, as one see through the boxplot, but did not produces strong biased estimates as GAMLSS. The same pattern is observed for

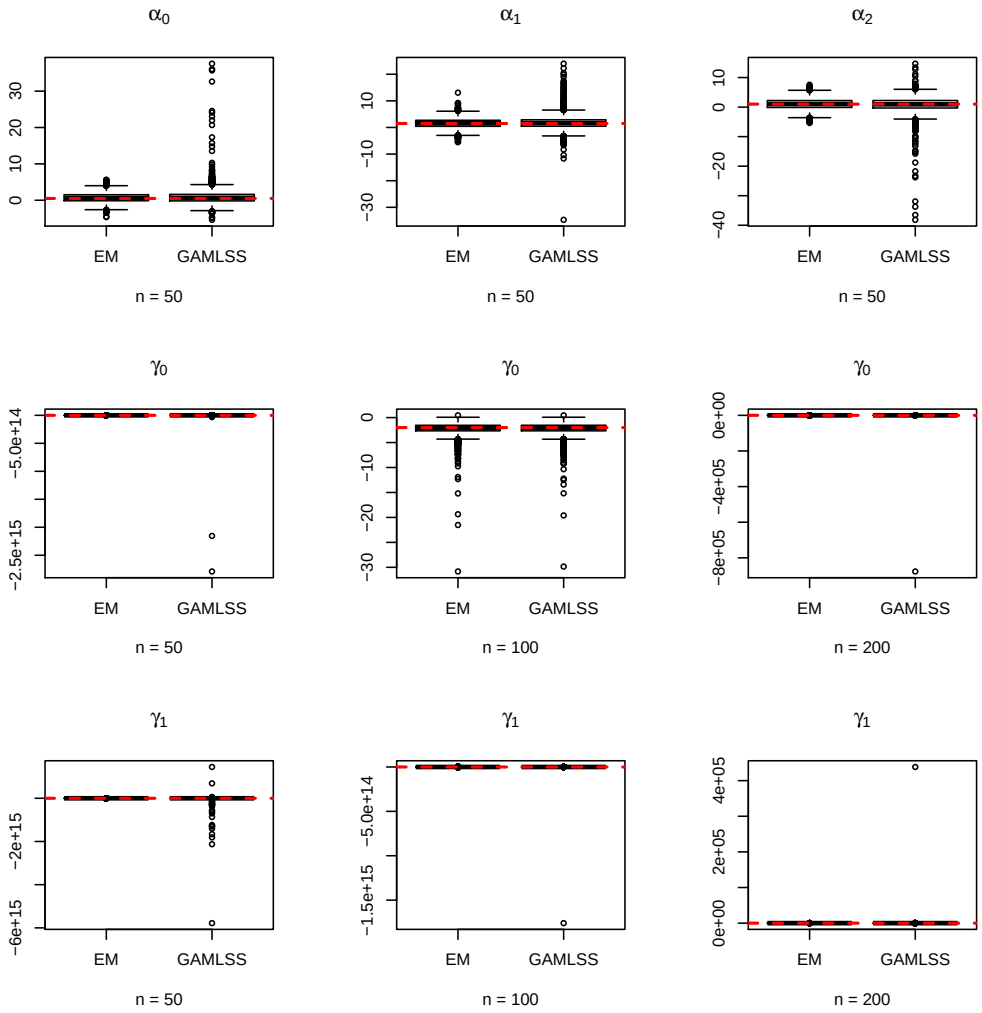


Figure 1: Estimates distribution of the ZINB model - 10% zero-inflation scenario

γ_1 parameters, but for samples of size 100 and 200.

An important remark is that, in a general way, the outliers of the γ's parameters from GAMLSS reach extremely high values. The dotted red line represents the real parameters value. Another important remark is that we decided do not put the box-plots of the β's parameters because of its good behavior for the EM algorithm and the GAMLSS estimates.

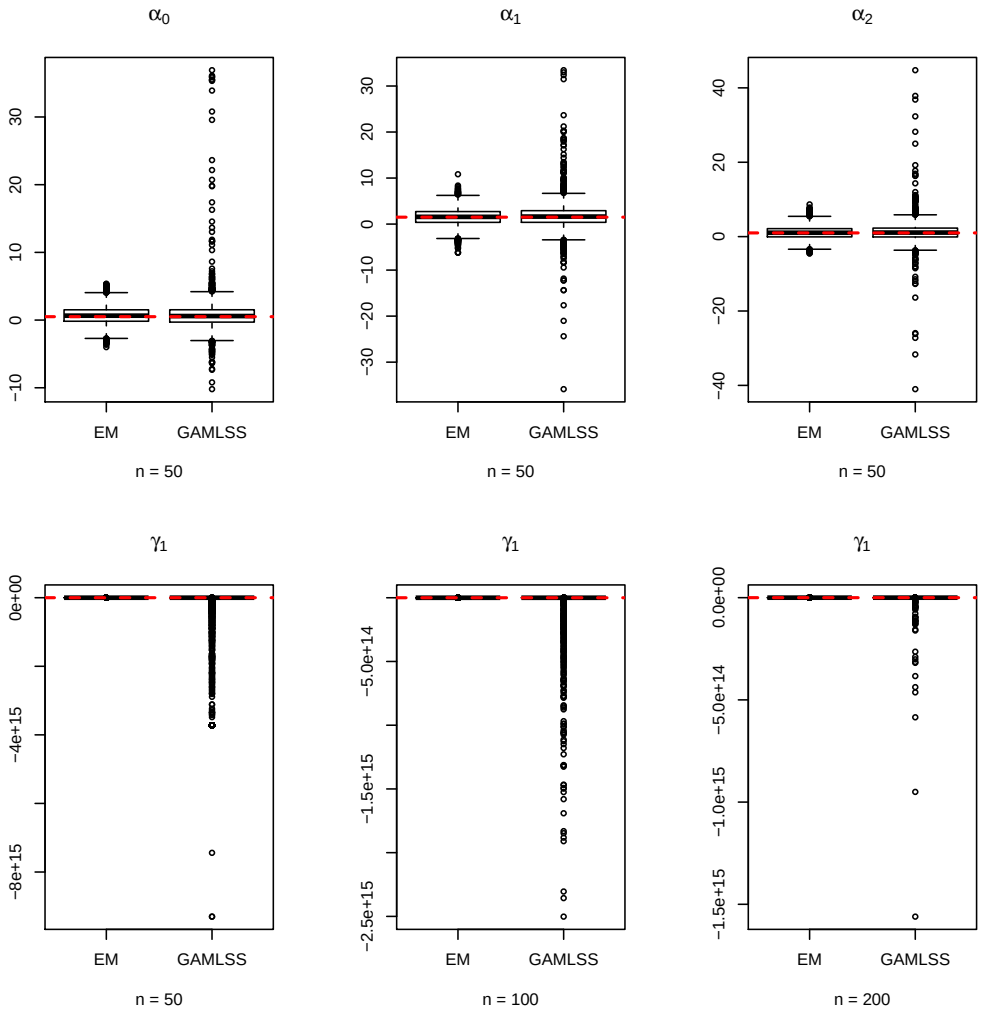


Figure 2: Estimates distribution of the ZINB model - 30% zero-inflation scenario

The assessment of figure 2 reveals that GAMLSS produced several outliers for the α's parameters, with a scale that goes from -40 to 40. Different from what happened in the 10% zero-inflation scenario, here one can realize that the bloxplots of γ_1, for any of the samples size, have lots of outliers that reach extreme high values, explaining the bad estimates presented in table 2, except for $n = 300$, whereas only one outlier is extremely high and if one remove that sample, then the estimate would be good. One more time, we reinforce that the EM algorithm could deal well with that issue and have presented good estimates, even for small samples size.

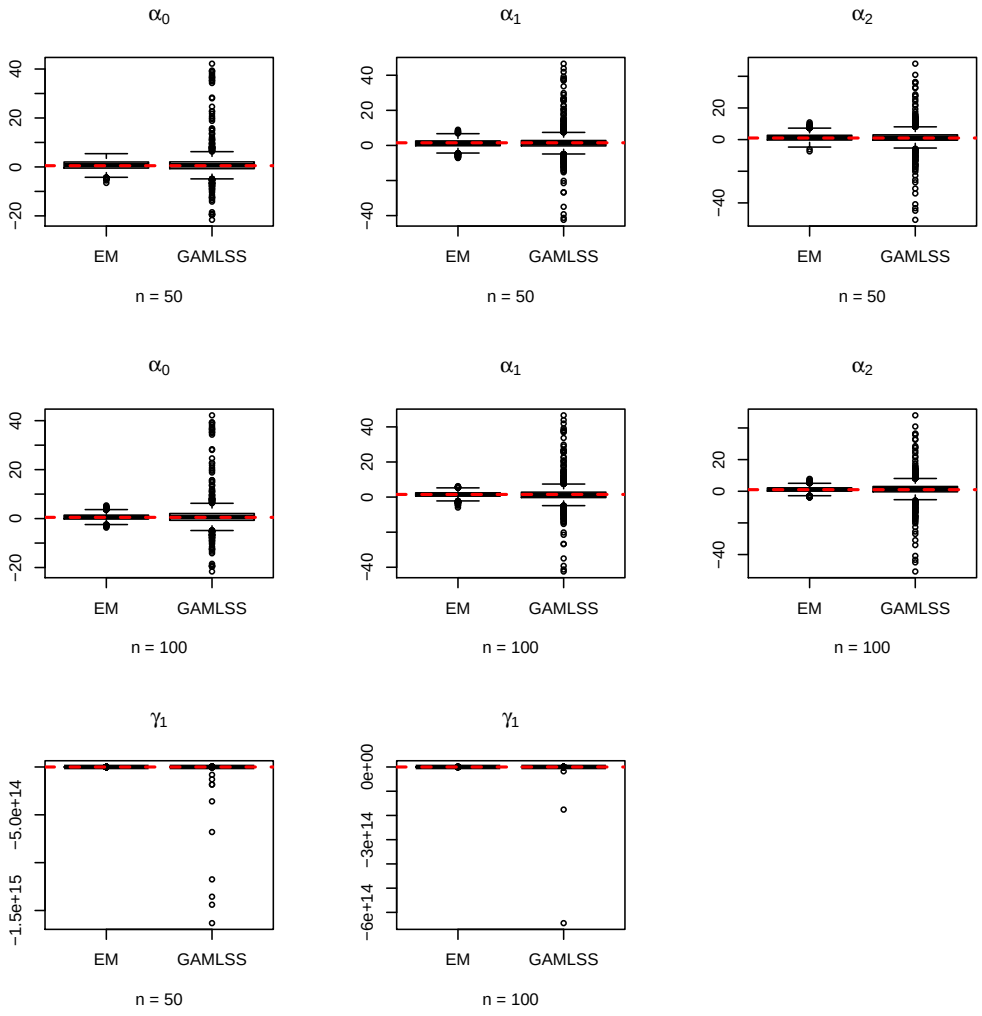


Figure 3: Estimates distribution of the ZINB model - 50% zero-inflation scenario

Figure 3 represents the estimates bloxplots of the 50% zero-inflation scenario. The analyzes of α's for 50 samples size are quite similar to the 10% and the 30% scenarios, with outliers produced by GAMLSS in a scale that goes to -40 to 40. The same pattern is also observed in 100 samples size, but it did not happen in the previous scenarios for samples of size 100. In this case, the many extreme values for γ_1 occur only for 50 samples size, once the bad estimate for the 100 samples sizes occur because only one extreme outlier, what could be solved removing that sample.

Essentially, GAMLSS use to produces hugely biased estimates for γ's, even for large samples size, as we noticed for γ_1 at the 30% zero-inflation scenario. Even in cases involving only one questionable sample, it is important to remember that such cases with extreme values produced did not arise from the EM algorithm estimates. Another important observation is that even GAMLSS has been produced not so biased estimates for the α's parameters and not so RMSE high values, we could observe the presence of lots of outliers that sometimes reach a large range, which can lead to unpleasant estimates for the dispersion parameter ϕ, as we will see at table 4.

Table 4: Average estimates of μ, ϕ and τ parameters from the ZINB model

	10% zero-inflation scenario			30% zero-inflation scenario			50% zero-inflation scenario		
$n = 50$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	9.508	6.693	0.134	11.338	7.063	0.276	11.902	6.406	0.489
$\overline{\Theta}_{\text{EM}}$	9.536	15.388	0.130	11.363	13.630	0.268	12.003	12.478	0.469
$\overline{\Theta}_{\text{GAMLSS}}$	9.542	1×10^{12}	NA	11.359	1×10^{13}	NA	12.010	2×10^{14}	0.465
$n = 100$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	10.506	6.316	0.130	11.499	6.847	0.259	10.377	5.800	0.520
$\overline{\Theta}_{\text{EM}}$	10.495	8.786	0.128	11.502	9.843	0.258	10.392	9.479	0.515
$\overline{\Theta}_{\text{GAMLSS}}$	10.499	12.281	0.128	11.506	10.434	NA	10.397	5×10^{11}	0.514
$n = 200$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	10.188	6.382	0.122	10.903	6.72	0.271	10.391	6.569	0.516
$\overline{\Theta}_{\text{EM}}$	10.196	7.490	0.120	10.895	8.059	0.270	10.385	8.600	0.515
$\overline{\Theta}_{\text{GAMLSS}}$	10.199	7.621	NA	10.898	8.213	0.270	10.391	9.180	0.515
$n = 300$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	10.657	6.470	0.127	11.582	6.712	0.266	10.875	6.130	0.499
$\overline{\Theta}_{\text{EM}}$	10.652	7.174	0.126	11.578	7.486	0.265	10.866	7.271	0.499
$\overline{\Theta}_{\text{GAMLSS}}$	10.656	7.280	0.126	11.583	7.580	0.265	10.872	7.388	0.499

By looking table 4, we observe good average estimates of μ and τ produced by the EM algorithm for the three scenarios. GAMLSS also produced good estimates, except for τ in the 10% zero-inflation scenario, on 50 and 200 samples size, and in the 30% zero-inflation scenario, on 100 and 200 samples size, whom could not produce an average estimate for τ because of the extreme γ's parameters estimates values. We highlight that for 200 samples size, at the 10% zero-inflation scenario, only one sample produced a high value for γ_0 and γ_1 and, removing that sample, probably GAMLSS would produces a good estimate. However, for the other samples size, several samples produced γ's estimates with huge values.

Despite GAMLSS did not produces strong biased values for α's parameters, its had several outliers that reached a range between -40 and 40, which reflected in the ϕ estimative, because it generated really biased estimates for 50 and 100 samples size, getting better only for 200 samples size or higher. It is true that the EM algorithm did not produces so positive estimates, but they were not so biased for samples of size 100 or higher. For samples of size 50 the estimates are biased, but not so strong as those produced by GAMLSS, and it is well known that modeling the dispersion parameter is not truly easy.

Now, the estimates of the ZIPIG regression model will be presented at table 5 for the three zero-inflation scenarios. A general analysis of table 5 shows that the estimates obtained through the EM algorithm performs well about bias and RMSE criteria, even for small samples size, except maybe for γ's parameters on samples of size 50 in the 10% zero-inflation scenario, specially γ_0 and γ_2, that are biased and with a high RMSE value.

Table 5: Mean and root of the mean square error, in parentheses, of the parameters estimates for the ZIPIG model

ZIPIG	10% zero-inflation scenario			
Θ	$n = 50$	$n = 100$	$n = 200$	$n = 300$
β_0	0.987 (0.368)	0.994 (0.246)	0.993 (0.179)	0.996 (0.143)
β_1	1.001 (0.497)	1.000 (0.249)	1.002 (0.197)	1.001 (0.171)
β_2	1.510 (0.429)	1.505 (0.303)	1.505 (0.209)	1.503 (0.164)
α_0	0.673 (3.121)	0.603 (1.365)	0.522 (0.813)	0.524 (0.653)
α_1	2.023 (3.867)	1.600 (1.814)	1.578 (1.121)	1.516 (0.850)
α_2	1.097 (3.544)	1.042 (1.730)	1.036 (1.105)	1.027 (0.838)
γ_0	−3.262 (10.037)	−2.282 (3.177)	−2.083 (0.850)	−2.061 (0.636)
γ_1	1.538 (4.688)	1.131 (1.475)	1.042 (0.399)	1.034 (0.403)
γ_2	−2.399 (20.419)	−1.591 (3.193)	−1.548 (1.385)	−1.530 (1.044)
ZIPIG	30% zero-inflation scenario			
Θ	$n = 50$	$n = 100$	$n = 200$	$n = 300$
β_0	0.987 (0.499)	0.994 (0.286)	0.997 (0.181)	0.994 (0.162)
β_1	1.005 (0.479)	0.998 (0.323)	0.998 (0.195)	1.001 (0.173)
β_2	1.506 (0.509)	1.504 (0.360)	1.503 (0.211)	1.505 (0.190)
α_0	0.691 (3.356)	0.600 (1.752)	0.539 (1.036)	0.531 (0.764)
α_1	2.158 (4.240)	1.743 (2.112)	1.573 (1.195)	1.558 (0.969)
α_2	1.093 (3.930)	1.031 (2.099)	1.030 (1.213)	1.010 (0.991)
γ_0	0.521 (1.118)	0.521 (0.686)	0.510 (0.502)	0.509 (0.418)
γ_1	−1.735 (2.024)	−1.793 (1.738)	−1.592 (0.691)	−1.564 (0.503)
γ_2	−2.121 (2.006)	−2.080 (1.383)	−2.042 (0.899)	−2.026 (0.751)
ZIPIG	50% zero-inflation scenario			
Θ	$n = 50$	$n = 100$	$n = 200$	$n = 300$
β_0	0.997 (0.588)	0.985 (0.384)	0.992 (0.251)	0.998 (0.193)
β_1	1.007 (0.478)	1.003 (0.359)	1.000 (0.253)	1.001 (0.196)
β_2	1.491 (0.722)	1.513 (0.429)	1.505 (0.292)	1.497 (0.222)
α_0	1.210 (5.246)	0.710 (2.146)	0.558 (1.249)	0.551 (0.979)
α_1	2.075 (6.113)	1.854 (2.752)	1.643 (1.672)	1.587 (1.204)
α_2	0.921 (6.552)	0.891 (2.695)	1.054 (1.649)	1.004 (1.236)
γ_0	1.003 (1.172)	1.036 (0.744)	1.021 (0.457)	1.009 (0.387)
γ_1	−0.665 (1.229)	−0.532 (0.450)	−0.521 (0.316)	−0.516 (0.255)
γ_2	−1.515 (1.691)	−1.555 (1.131)	−1.531 (0.788)	−1.508 (0.612)

The comparison between the EM algorithm estimates and GAMLSS estimates has not been made because GAMLSS could not fit most of the 4500 samples of the Monte Carlo study. Then, we present at table below how many samples GAMLSS could fit among the 4500 samples, considering the three scenarios of different zero-inflation rates and for samples of size 50, 100, 200 and 300.

Table 6: Number of samples fitted by GAMLSS in 4500 samples

Scenarios	$n = 50$	$n = 100$	$n = 200$	$n = 300$
10% zero-inflation	1239	3857	3975	1
30% zero-inflation	208	11	1	88
50% zero-inflation	5	0	0	1098

The figure below allow us to check how the parameters estimates of some of the 4500 samples are distributed for the ZIPIG case, making a comparison between the estimates of the EM algorithm and GAMLSS for the 10% zero-inflation scenario, the only one that GAMLSS could fit more than 1000 samples, considering samples of size 50, 100 and 200.

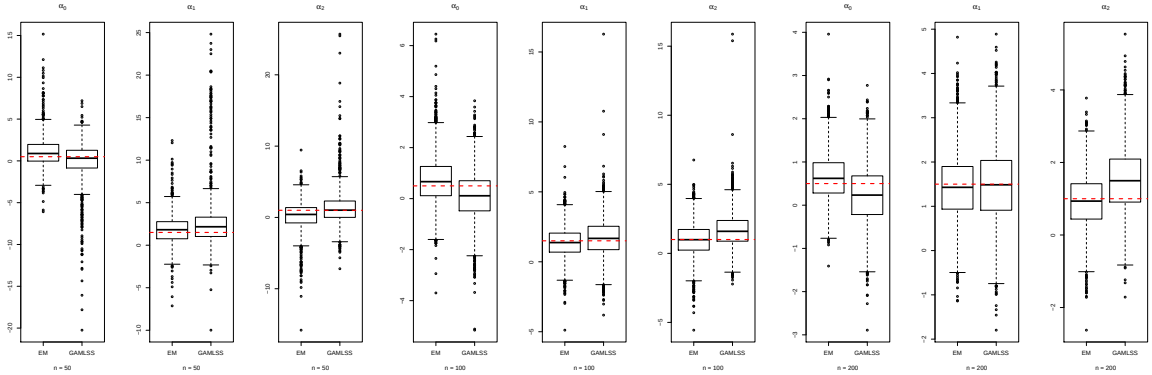


Figure 4: Estimates distribution of the ZIPIG model - 10% zero-inflation scenario

We are drawing attention to the α's parameters because the estimates produced by GAMLSS are underestimated for α_0 in samples of size 100 and 200 and overestimated for α_2, which can conduct to bad estimates of the dispersion parameter ϕ.

We now observe tables 7 to 9, that hold the empirical standard deviation of the parameters estimates of the model and theoretical standard deviation of the parameters for the ZINB regression model, it is, the average of the standard errors extracted from the observed information matrix. The tables reveal good results even for 50 samples size, leading to the conclusion that the standard errors obtained through the observed information matrix are appropriate for estimating the parameter estimators standard errors.

Table 7: Empirical and theoretical standard errors of the estimates for the parameters of the ZINB model - 10% zero-inflation scenario

10% zero-inflation	β_0	β_1	β_2	α_0	α_1	α_2	γ_0	γ_1	γ_2
$n = 50$									
Empirical	0.274	0.370	0.320	1.313	1.752	1.752	4.684	1.899	7.384
Theoretical	0.263	0.351	0.309	1.347	1.909	1.932	5.241	2.569	3.868
$n = 100$									
Empirical	0.194	0.218	0.221	0.819	1.023	1.109	1.169	0.545	1.774
Theoretical	0.189	0.211	0.215	0.796	1.024	1.097	0.876	0.442	1.490
$n = 200$									
Empirical	0.122	0.141	0.140	0.525	0.702	0.705	0.571	0.321	0.962
Theoretical	0.118	0.138	0.136	0.508	0.698	0.688	0.526	0.296	0.948
$n = 300$									
Empirical	0.095	0.110	0.111	0.421	0.564	0.549	0.431	0.242	0.737
Theoretical	0.098	0.110	0.113	0.416	0.561	0.548	0.423	0.234	0.710

Table 8: Empirical and theoretical standard errors of the estimates for the parameters of the ZINB model - 30% zero-inflation scenario

30% zero-inflation	β_0	β_1	β_2	α_0	α_1	α_2	γ_0	γ_1	γ_2
$n = 50$									
Empirical	0.352	0.317	0.366	1.338	1.809	1.728	0.889	2.894	1.484
Theoretical	0.337	0.305	0.346	1.391	1.988	1.931	0.816	12.338	1.352
$n = 100$									
Empirical	0.210	0.205	0.240	0.937	1.137	1.245	0.615	1.179	1.037
Theoretical	0.204	0.196	0.235	0.949	1.184	1.262	0.597	1.329	1.001
$n = 200$									
Empirical	0.143	0.153	0.159	0.620	0.760	0.756	0.373	0.502	0.634
Theoretical	0.140	0.149	0.155	0.626	0.763	0.773	0.376	0.460	0.633
$n = 300$									
Empirical	0.116	0.121	0.130	0.482	0.591	0.630	0.309	0.339	0.543
Theoretical	0.113	0.119	0.129	0.478	0.588	0.633	0.307	0.338	0.540

Table 9: Empirical and theoretical standard errors of the estimates for the parameters of the ZINB model - 50% zero-inflation scenario

50% zero-inflation	β_0	β_1	β_2	α_0	α_1	α_2	γ_0	γ_1	γ_2
$n = 50$									
Empirical	0.450	0.418	0.529	1.863	2.061	2.295	0.942	0.671	1.424
Theoretical	0.429	0.397	0.509	2.327	2.635	2.855	0.923	0.607	1.393
$n = 100$									
Empirical	0.261	0.288	0.316	1.165	1.439	1.519	0.490	0.373	0.887
Theoretical	0.253	0.269	0.307	1.278	1.633	1.744	0.490	0.362	0.883
$n = 200$									
Empirical	0.170	0.189	0.206	0.800	0.942	1.045	0.327	0.246	0.546
Theoretical	0.169	0.186	0.201	0.792	0.970	1.036	0.326	0.241	0.542
$n = 300$									
Empirical	0.145	0.159	0.165	0.645	0.784	0.792	0.277	0.188	0.440
Theoretical	0.143	0.154	0.162	0.643	0.801	0.785	0.271	0.183	0.433

Tables 10 to 12 hold the empirical standard deviation of the parameters estimates of the model and theoretical standard deviation of the parameters for the ZIPIG regression model. The tables lead to the same conclusion as in the ZINB case, it is, reveal good results even for samples of size 50, except for γ's in the 10% zero-inflation scenario, leading to the conclusion that the standard errors obtained through the observed information matrix are appropriate for estimating the parameter estimators standard errors.

Table 10: Empirical and theoretical standard errors of the estimates for the parameters of the ZIPIG model - 10% zero-inflation scenario

10% zero-inflation	β_0	β_1	β_2	α_0	α_1	α_2	γ_0	γ_1	γ_2
$n = 50$									
Empirical	0.260	0.351	0.303	2.203	2.710	2.505	7.041	3.293	14.425
Theoretical	0.242	0.317	0.282	1.941	2.348	2.214	500.479	254.738	6.229
$n = 100$									
Empirical	0.174	0.176	0.215	0.962	1.281	1.223	2.238	1.039	2.257
Theoretical	0.171	0.170	0.208	0.934	1.197	1.200	3.364	1.705	1.581
$n = 200$									
Empirical	0.126	0.140	0.148	0.575	0.791	0.781	0.598	0.280	0.979
Theoretical	0.126	0.136	0.148	0.577	0.784	0.788	0.576	0.269	0.958
$n = 300$									
Empirical	0.101	0.121	0.116	0.462	0.601	0.592	0.448	0.284	0.738
Theoretical	0.098	0.116	0.115	0.468	0.606	0.599	0.451	0.283	0.729

Table 11: Empirical and theoretical standard errors of the estimates for the parameters of the ZIPIG model - 30% zero-inflation scenario

30% zero-inflation	β_0	β_1	β_2	α_0	α_1	α_2	γ_0	γ_1	γ_2
$n = 50$									
Empirical	0.353	0.339	0.360	2.369	2.962	2.779	0.791	1.422	1.416
Theoretical	0.322	0.313	0.327	2.077	2.513	2.411	0.745	2.225	1.310
$n = 100$									
Empirical	0.202	0.228	0.254	1.237	1.484	1.484	0.485	1.212	0.976
Theoretical	0.191	0.216	0.239	1.195	1.442	1.461	0.479	1.376	0.959
$n = 200$									
Empirical	0.128	0.138	0.149	0.732	0.843	0.858	0.355	0.484	0.635
Theoretical	0.127	0.132	0.151	0.733	0.862	0.856	0.349	0.438	0.625
$n = 300$									
Empirical	0.114	0.123	0.135	0.539	0.684	0.701	0.295	0.353	0.531
Theoretical	0.112	0.121	0.135	0.539	0.685	0.701	0.295	0.339	0.528

Table 12: Empirical and theoretical standard errors of the estimates for the parameters of the ZIPIG model - 50% zero-inflation scenario

50% zero-inflation	β_0	β_1	β_2	α_0	α_1	α_2	γ_0	γ_1	γ_2
$n = 50$									
Empirical	0.416	0.338	0.511	3.676	4.304	4.633	0.829	0.861	1.195
Theoretical	0.357	0.292	0.432	3.555	3.695	4.181	0.767	0.741	1.116
$n = 100$									
Empirical	0.271	0.254	0.303	1.510	1.930	1.904	0.525	0.317	0.799
Theoretical	0.255	0.240	0.286	1.387	1.808	1.806	0.508	0.300	0.768
$n = 200$									
Empirical	0.177	0.179	0.207	0.882	1.178	1.166	0.323	0.223	0.557
Theoretical	0.172	0.172	0.202	0.856	1.145	1.143	0.320	0.219	0.554
$n = 300$									
Empirical	0.136	0.139	0.157	0.691	0.849	0.874	0.274	0.180	0.433
Theoretical	0.134	0.137	0.155	0.683	0.840	0.873	0.274	0.179	0.435

We have also performed a simulation study considering the zero-inflation parameter τ near to zero, it is, when the data set does not have zeros excess. When one takes a look at the γ's parameters estimates, they are far from the real values especially for samples of size 50 and 100, being closer to the real parameters for samples of size 200 or higher. However, when we take a look at the average estimates of μ, ϕ and τ parameters, we may notice a satisfactory performance, as showed by table 13 below

Table 13: Average estimates of μ, ϕ and τ parameters from the ZINB and the ZIPIG models

Non zero-inflation scenario							
ZINB				ZIPIG			
$n = 50$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$n = 50$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	11.857	6.495	0.017	$\overline{\Theta}$	11.362	6.620	0.015
$\overline{\Theta}_{\text{EM}}$	11.902	84.881	0.021	$\overline{\Theta}_{\text{EM}}$	11.409	1021.983	0.015
$n = 100$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$n = 100$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	10.647	6.801	0.016	$\overline{\Theta}$	11.548	7.016	0.014
$\overline{\Theta}_{\text{EM}}$	10.652	9.252	0.019	$\overline{\Theta}_{\text{EM}}$	11.557	10.800	0.017
$n = 200$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$n = 200$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	10.822	6.396	0.015	$\overline{\Theta}$	11.325	6.368	0.015
$\overline{\Theta}_{\text{EM}}$	10.823	7.324	0.016	$\overline{\Theta}_{\text{EM}}$	11.339	7.436	0.018
$n = 300$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$	$n = 300$	$\overline{\mu}$	$\overline{\phi}$	$\overline{\tau}$
$\overline{\Theta}$	10.756	6.476	0.016	$\overline{\Theta}$	10.568	6.287	0.015
$\overline{\Theta}_{\text{EM}}$	10.760	6.981	0.016	$\overline{\Theta}_{\text{EM}}$	10.566	6.902	0.015

5 Empirical Illustration

This chapter is dedicated to show how useful the proposed class is through an application to a real data set, which is described as follows. The data, reported by Ridout et al. (1998) and also analyzed, for instance, by Ridout et al. (2001) and Garay et al. (2011), is an apple cultivar data with the number of roots produced by 270 micropropagated shoots of the columnar apple cultivar Trajan.

According to Ridout et al. (1998), during the rooting period, all shoots were maintained under identical conditions, but the shoots themselves were cultured on media containing one of four different concentrations of the cytokinin BAP, in growth cabinets with an 8 or 16 hour photoperiod. There were 30 or 40 shoots of each of these eight treatment combinations. Of the 140 shoots produced under the 8 hour photoperiod, only 2 failed to produce roots, but 62 of the 130 shoots produced under the 16 hour photoperiod failed to root. Thus, the response variable, denoted by w_i, and the covariates of interest are

- w_i : count of roots,
- x_{i1} : concentrations of the cytokinin BAP,
- x_{i2} : photoperiod ($0 = 8$ hrs, $1 = 16$ hrs).

Therefore, the model may be write as follows

- $\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$,
- $\log(\phi_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i}$,
- $\text{logit}(\tau_i) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}$,

for $i = 1, \ldots, 270$. Two scenarios will be taking into account for fit the apple cultivar data, it is, two models with that regression structure will be considered, the first one is the zero-inflated negative binomial regression model and the second one is the zero-inflated Poisson-inverse Gaussian regression model.

Starting our study by the zero-inflated negative binomial case, the estimates of the full fitted model is presented at table 14, as well as its standard errors, z values and p values.

Table 14: Estimates, standard errors, z values and p values of the full ZINB model fit

Θ	Est.	SE	z value	p value
β_0	1.961	0.070	27.835	0.000
β_1	0.003	0.006	0.450	0.653
β_2	-0.383	0.092	-4.140	0.000
α_0	1.641	0.634	2.588	0.010
α_1	0.187	0.098	1.898	0.058
α_2	-1.522	0.689	-2.210	0.027
γ_0	-4.423	0.795	-5.560	0.000
γ_1	0.016	0.031	0.502	0.615
γ_2	4.078	0.773	5.278	0.000

One can notice through table 14 that the covariate concentrations of the cytokinin BAP, x_1, is not significant for modeling both the mean μ_i and the zero-inflation parameter τ_i of the count of roots, once the p values related to the parameters β_1 and γ_1 are not significant at the 5% usual significance level. But, in a first moment, that covariate seems to be considerable for explain the dispersion parameter, since its value is closer to the significance level. For this reason, a second and reduced model was fitted, considering $\beta_1 = \gamma_1 = 0$, as it is possible to see at table 15.

Table 15: Estimates, standard errors, z values and p values of the reduced ZINB model fit

Θ	Est.	SE	z value	p value
β_0	1.987	0.038	52.758	0.000
β_2	−0.371	0.087	−4.262	0.000
α_0	1.613	0.633	2.546	0.011
α_1	0.187	0.101	1.860	0.063
α_2	−1.376	0.635	−2.166	0.030
γ_0	−4.296	0.756	−5.680	0.000
γ_2	4.114	0.780	5.272	0.000

After fitting the reduced model, at the 5% significance level, table 15 suggests that the parameter α_1 is not significant indeed. Therefore, how the covariate x_1 is showing to be not relevant, a third model is presented in the following table, taking into account just the covariate photoperiod, x_2.

Table 16: Estimates, standard errors, z values and p values of the final reduced ZINB model fit

Θ	Est.	SE	z value	p value
β_0	1.972	0.037	53.618	0.000
β_2	−0.303	0.088	−3.425	0.001
α_0	3.083	0.522	5.903	0.000
α_2	−1.562	0.660	−2.369	0.018
γ_0	−4.387	0.831	−5.277	0.000
γ_2	4.231	0.852	4.968	0.000

Now, taking a look at figure 5, which presents the simulated envelopes of the Pearson residual against the theoretical quantiles of the standard normal distribution, we may conclude that the fitting looks suitable, since the residuals remain all inside of the envelope and, furthermore, is difficult detect the presence of outliers.

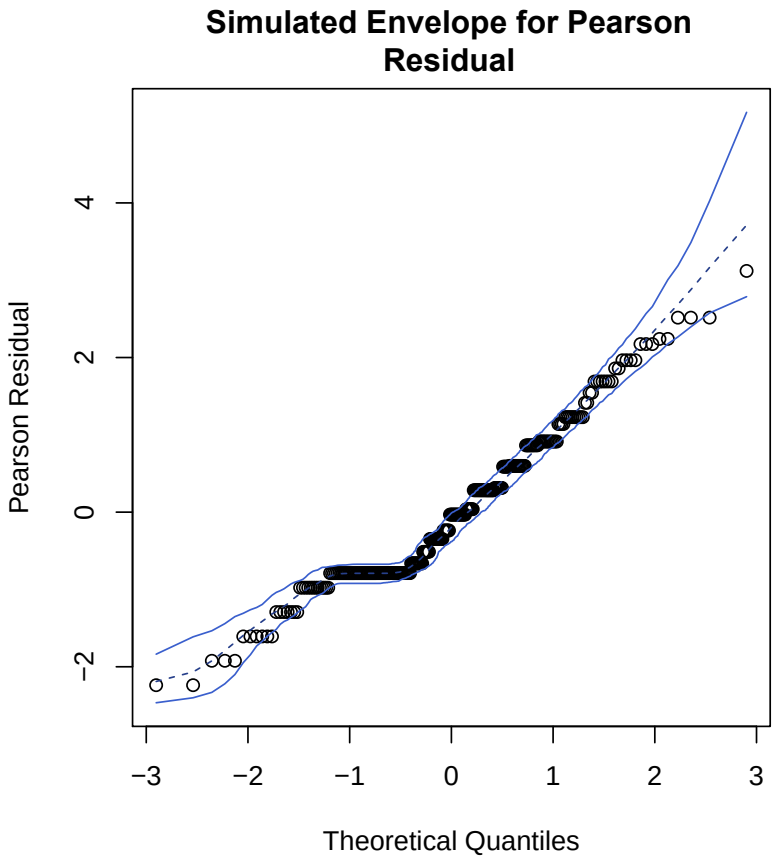


Figure 5: Simulated envelopes for the Pearson residual in the ZINB model

To verify the presence of observations that may be influential, lets take a look at figure 6 that contains the generalized Cook's distance measure. In the figure there are four plots, a general plot and one for each group of parameters, it is, for β's, α's and γ's. In essence, the plots indicate two observations as potential outliers that can pursue influence, observations 101 and 102. In order to identify how this observations can influence the fit, after the influence analysis of the ZIPIG model, we present a new fit for ZINB and ZIPIG regression models, but removing those points.

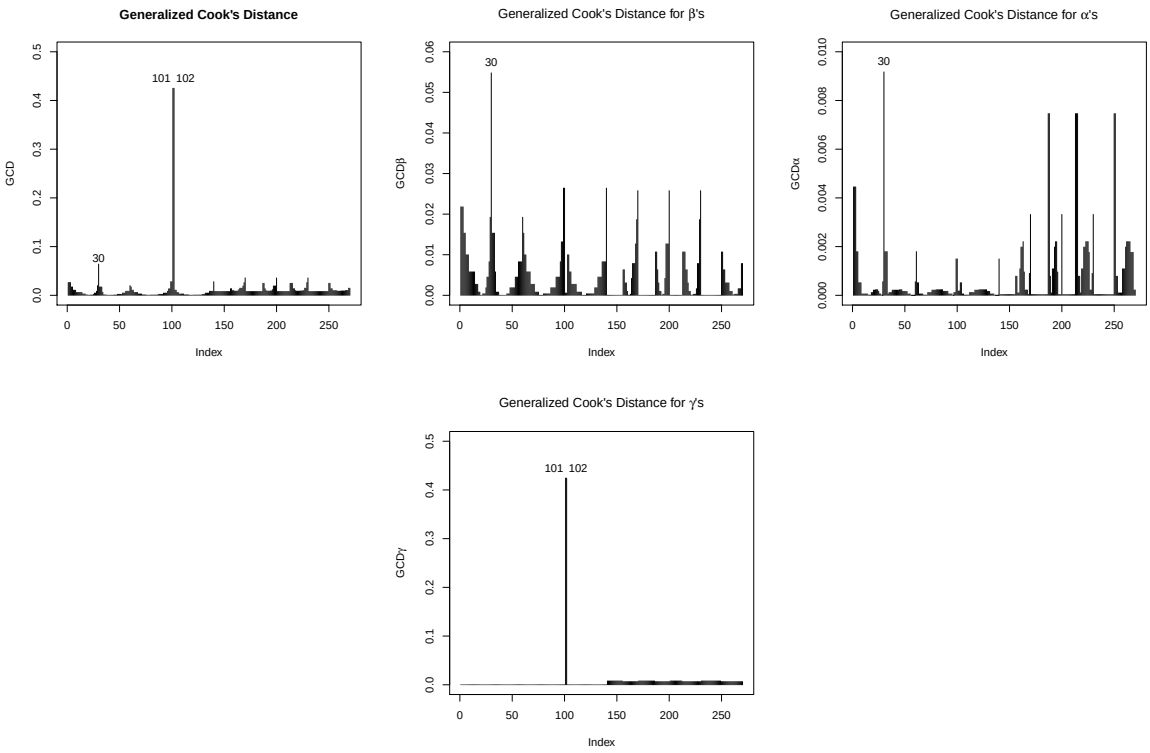


Figure 6: Generalized Cook's Distance of the ZINB model

Pursuing with the models fit, we now focus on zero-inflated Poisson-inverse Gaussian regression model. Thus, the estimates of the full fitted model is presented at table 17, as well as its standard errors, z values and p values.

Table 17: Estimates, standard errors, z values and p values of the full ZIPIG model fit

Θ	Est.	SE	z value	p value
β_0	1.970	0.070	28.161	0.000
β_1	0.002	0.006	0.299	0.765
β_2	-0.358	0.084	-4.255	0.000
α_0	1.657	0.748	2.215	0.027
α_1	0.184	0.111	1.662	0.097
α_2	-1.275	0.701	-1.819	0.069
γ_0	-4.363	0.790	-5.524	0.000
γ_1	0.009	0.029	0.303	0.762
γ_2	4.133	0.771	5.360	0.000

Analyzing table 17, it is possible to reach the same conclusions as that pointed about table 14, wherein the covariate x_1 is not significant for modeling the mean, the zero-inflation and also the dispersion parameters of the count of roots, since the p values related to the parameters β_1, α_1 and γ_1 are not significant at the 5% usual significance level. The covariate x_2 seems significant, except maybe for the α_2, related to the dispersion parameter. However, the p value for α_2 is smaller than the 10% level and one can consider keep the covariate and decides or not by its exclusion after fit the reduced model. For this reason, a second and reduced model was fitted and the covariate x_2 has been kept, considering $\beta_1 = \alpha_1 = \gamma_1 = 0$.

Table 18: Estimates, standard errors, z values and p values of the reduced ZIPIG model fit

Θ	Est.	SE	z value	p value
β_0	1.972	0.037	53.753	0.000
β_2	−0.295	0.086	−3.432	0.001
α_0	3.103	0.526	5.894	0.000
α_2	−1.523	0.682	−2.231	0.026
γ_0	−4.373	0.819	−5.339	0.000
γ_2	4.235	0.839	5.047	0.000

The analysis of the reduced model at table 18 suggests that we have found a final model, since all parameters are significant at 5% significance level. Therefore, the final proposed model for both the ZINB and the ZIPIG regression models agree one each other because both reach the same conclusion, it is, that only the covariate photoperiod was significant for modeling the count of apple roots. An interesting remark is that the ZINB model reaches the final reduced model in three steps whereas the ZIPIG model reaches the final reduced model directly, in two steps.

Now, taking a look at figure 7, which presents the simulated envelopes of the Pearson residual against the theoretical quantiles of the standard normal distribution, we may conclude that the fitting looks suitable, since the residuals remain almost all inside of the envelope, except perhaps the one next to the last.

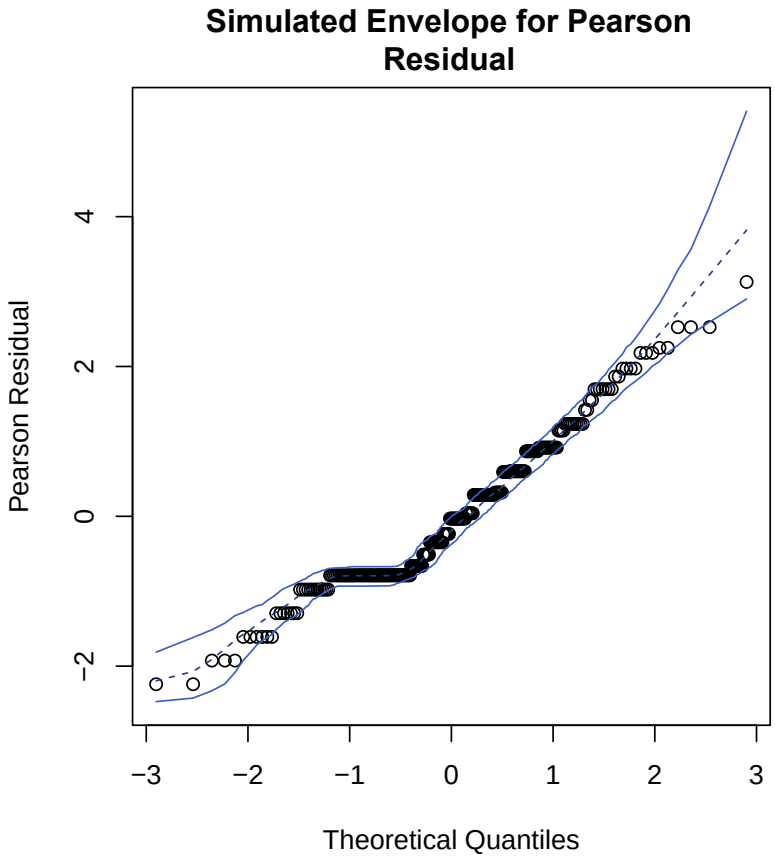


Figure 7: Simulated envelopes for the Pearson residual in the ZIPIG model

To verify the presence of observations that may be influential in the ZIPIG regression model, we should look at figure 8 that contains the generalized Cook's distance measure for its case. The plots indicate the same two observations as that pointed in the ZINB regression model as potential outliers that can pursue influence. Then, we present in what follows the results of the fit for the ZINB and the ZIPIG regression models case after removing 101 and 102 observations.

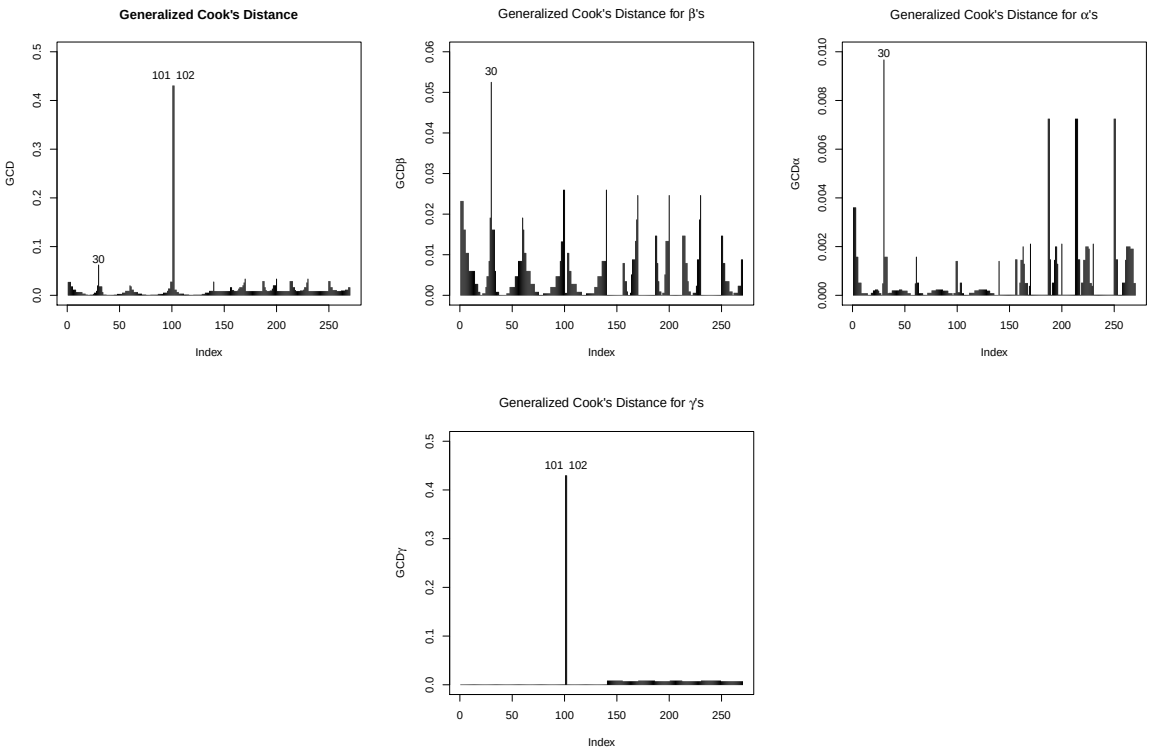


Figure 8: Generalized Cook's Distance of the ZIPIG model

Table 19: Estimates, standard errors, z values and p values after removing 101 and 102 observations

ZINB	Θ	Est.	z value	p value
β_0	1.974	0.036	54.306	0.000
β_2	-0.305	0.088	-3.455	0.001
α_0	3.134	0.527	5.951	0.000
α_2	-1.612	0.663	-2.429	0.015
γ_0	-11.581	27.868	-0.416	0.678
γ_2	11.425	27.869	0.410	0.682
ZIPIG	Θ	Est.	z value	p value
β_0	1.974	0.036	54.714	0.000
β_2	-0.297	0.086	-3.460	0.001
α_0	3.197	0.553	5.785	0.000
α_2	-1.621	0.702	-2.310	0.021
γ_0	-10.449	15.847	-0.659	0.510
γ_2	10.310	15.848	0.651	0.515

Table 19 provides a first print that ZIMP regression models perhaps are not the best choice to fit the chosen data set, once all γ's parameters are not significant at any usual significance level, but we should be careful in our assessment. Here we could notice the importance of the diagnostic analysis, that raise awareness of that maybe the data set is not zero-inflated.

Then, proceeding with a particular data set examination, one may notice that the data set can be divided in two data sets, the first one composed by those 140 shoots produced under the 8 hour photoperiod, wherein only 2 failed to produce roots, and the second one composed by 130 shoots produced under the 16 hour photoperiod, wherein 62 failed to produce roots. It is important to realize that the observations pointed as possible influential are exactly the only 2 shoots that failed under the 8 hour photoperiod.

Thus, the choice of the apple cultivar data set was interesting because of some reasons, such as researches must pay attention to their data sets, as well as which model should be applied, once the fit can be compromised. Furthermore, one can realize the importance of the diagnostics analysis. We must acknowledge that we have made just the global influence analysis, but this was enough to identify that maybe the data set should not be directly be fitted by a ZIMP regression model. The fact is the data is composed by to distinct groups, one non-zero-inflated and a second one that have zeros excess. Therefore, the reader will find in the next section the fit of the apple cultivar data set, but now divided in two data sets, the first one with no zero excess and the second one zero-inflated.

We highlight to the fact that some analyzes made by Garay et al. (2011) are compromised, especially about the diagnostic analysis on the application section, unfortunately because of a error type in the data set, once they reported one of the concentrations of the cytokinin BAP as 18.6 instead of 17.6. Despite they have found similar parameters

estimates, they have found different observations as possible influential points through the GCD measure, observations 191 and 192, while we have found observations 101 and 102, changing completely assessments about that data set.

We would like to draw attention to the fact that we fit GAMLSS for the ZINB and the ZIPIG cases. As expected, the ZINB case was fitted successfully, but for the ZIPIG case, as already indicated in the simulation chapter, GAMLSS could not fit this data set, reporting the following error: *"Error in RS(): The global deviance is increasing. Try different steps for the parameters or the model maybe inappropriate".*

Lets check now the fit of the models without the zero-inflation structure, it is, the fit of the negative binomial regression model and the Poisson-inverse Gaussian regression model, as well as, the ZIP regression model in order to show its inadequacy. For that, we present in what follows a table with the observed versus the predicted values for the percentage of counts of apple roots equal to w for each model, but just for those shoots that were produced under the 16 hour photoperiod, according to the previous discussion made after de diagnostic analysis.

The analysis of the table 20 allows us to realize that the ZINB and the ZIPIG are the regression models that have had the better performance, since the difference between the observed percentage of counts values and the predicted ones are smaller than NB, PIG and ZIP regression models, especially for the percentage of zeros. While the ZINB and the ZIPIG regression models provide correctly predictions for the zeros count, the NB and the PIG models underestimate the zeros count. The bad performance occurs also at one and two roots counts, where the ZINB and the ZIPIG produces the best predictions, while the ZIP regression model underestimates and, on the other hand, the NB and the PIG regression models strongly overestimate.

Table 20: Observed x Predicted values of the apple roots count

16 hrs photoperiod		NB		PIG		ZIP		ZINB		ZIPIG	
w	Observed	Predic	Diff	Predic	Diff	Predic	Diff	Predic	Diff	Predic	Diff
0	47.70	43.00	−4.70	37.30	−10.40	47.70	0.00	47.70	0.00	47.70	0.00
1	5.40	15.20	9.80	22.20	16.80	1.20	−4.20	3.90	−1.50	3.50	−1.90
2	5.40	9.40	4.00	11.90	6.50	3.40	−2.00	5.80	0.40	5.80	0.40
3	6.20	6.60	0.40	7.00	0.80	6.10	−0.10	6.90	0.70	7.10	0.90
4	6.20	4.90	−1.30	4.60	−1.60	8.30	2.10	7.00	0.80	7.30	1.10
5	4.60	3.80	−0.80	3.20	−1.40	9.00	4.40	6.40	1.80	6.70	2.10
6	7.70	3.00	−4.70	2.30	−5.40	8.20	0.50	5.50	−2.20	5.60	−2.10
7	3.10	2.40	−0.70	1.80	−1.30	6.40	3.30	4.50	1.40	4.50	1.40
8	1.50	1.90	0.40	1.40	−0.10	4.30	2.80	3.50	2.00	3.40	1.90
9	5.40	1.60	−3.80	1.10	−4.30	2.60	−2.80	2.60	−2.80	2.50	−2.90
10	3.10	1.30	−1.80	0.90	−2.20	1.40	−1.70	1.90	−1.20	1.80	−1.30
11	1.50	1.10	−0.40	0.80	−0.70	0.70	−0.80	1.40	−0.10	1.30	−0.20
12	2.30	0.90	−1.40	0.60	−1.70	0.30	−2.00	0.90	−1.40	0.90	−1.40
RSDS		13.39		22.26		8.93		5.37		5.68	

(a) NB model

(b) PIG model

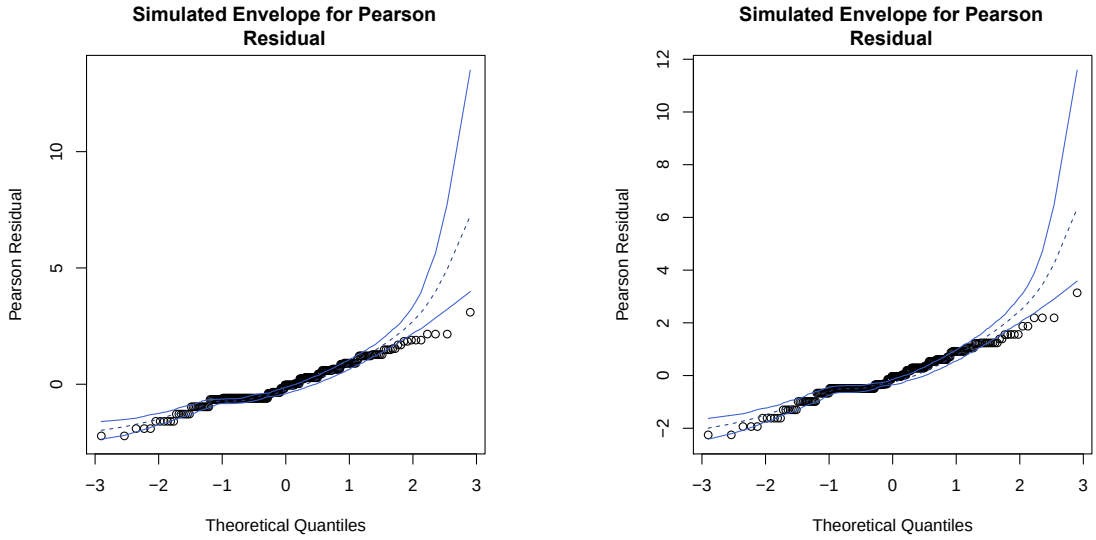


Figure 9: Simulated envelopes for the Pearson residual in the NB and the PIG regression models

In summary, inadequacy of the NB, PIG and the ZIP regression models may be noticed by the root of square difference sum (RSDS) measure, wherein the prediction error was higher than the prediction error of the ZINB and the ZIPIG regression models.

The results presented previously lead to the conclusion that the overdispersion might be caused from more than one source and one of them is the excess of zeros. Then, because of the heterogeneity of the data set that produces overdispersion, we may conclude that it is a scenario where a model that just carry out overdispersion, but that do not take into account the excess of zero, such the NB and the PIG regression models, can not handle. Despite the ZIP regression models handle the excess of zeros, we might remember that the Poisson distribution is equidispersed, reason why it showed poor fit if compared with the ZINB and the ZIPIG models.

The figures above strengthen our argument, once we plot the simulated envelopes of the Pearson residual against the theoretical quantiles of the standard normal distribution of the NB and the PIG regression models, wherein there are many residuals lying out of the envelopes in both models, indicating a possibly model misspecification.

5.1 Splitting Apple Cultivar Data Set

The figure 10 represents the frequency of the roots count, wherein the first one is the full data set, the second one is the data set considering the 8 hours of photoperiod and the last one is the data under the 16 hours of photoperiod. By looking figure 10 one can think the data is zero-inflated but, after dividing the data, it is possible to see the excess of zeros only under the 16 hours photoperiod case.

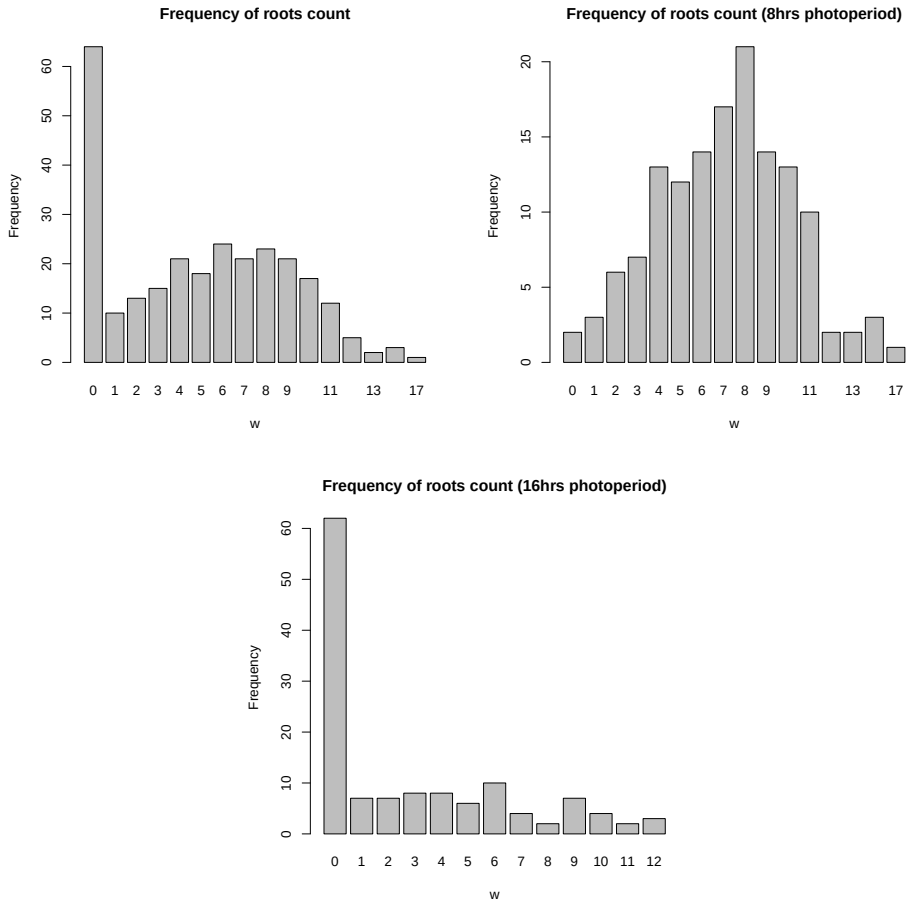


Figure 10: Frequency of roots count

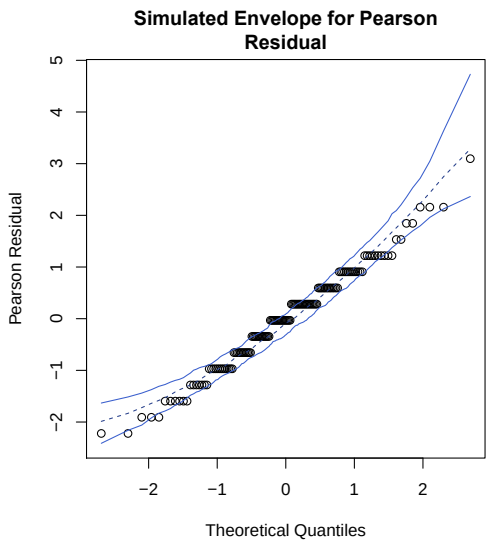
We tried to readjust the models for the divided data sets considering the covariate concentrations of the cytokinin BAP, x_1, but, as expected, it keeps insignificant and, for this reason, the following results are from a fit without covariates.

For the observed versus predicted values, we can realize that the NB and the PIG regression models fitted better for the first data set, under the 8 hour photoperiod, as already expected. But it is enjoyable that the ZINB and the ZIPIG regression models fitted as well as the NB and the PIG regression models, conclusion that arises from table 21, wherein the root of square difference sum is quite similar for all models, just as well from figure 11 wherein the residuals remain almost all inside the envelopes for all fitted models.

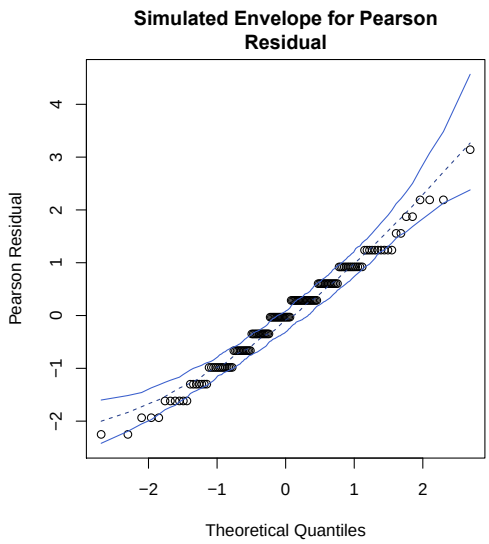
Table 21: Observed x Predicted values of the apple roots count under 8 hours photoperiod

8 hrs photoperiod		NB		PIG		ZIP		ZINB		ZIPIG	
w	Observed	Predic	Diff	Predic	Diff	Predic	Diff	Predic	Diff	Predic	Diff
0	1.40	0.30	−1.10	0.20	−1.20	1.40	0.00	1.40	0.00	1.40	0.00
1	2.10	1.40	−0.70	1.20	−0.90	0.50	−1.60	1.10	−1.00	1.00	−1.10
2	4.30	3.60	−0.70	3.40	−0.90	1.90	−2.40	3.00	−1.30	2.90	−1.40
3	5.00	6.60	1.60	6.50	1.50	4.60	−0.40	6.00	1.00	5.90	0.90
4	9.30	9.70	0.40	9.70	0.40	8.30	−1.00	9.10	−0.20	9.10	−0.20
5	8.60	11.90	3.30	12.10	3.50	11.90	3.30	11.70	3.10	11.80	3.20
6	10.00	12.90	2.90	13.10	3.10	14.30	4.30	12.90	2.90	13.10	3.10
7	12.10	12.40	0.30	12.70	0.60	14.70	2.60	12.70	0.60	12.90	0.80
8	15.00	11.00	−4.00	11.10	−3.90	13.20	−1.80	11.40	−3.60	11.50	−3.50
9	10.00	9.00	−1.00	9.00	−1.00	10.50	0.50	9.40	−0.60	9.40	−0.60
10	9.30	6.90	−2.40	6.90	−2.40	7.60	−1.70	7.10	−2.20	7.10	−2.20
11	7.10	5.00	−2.10	4.90	−2.20	5.00	−2.10	5.10	−2.00	5.10	−2.00
12	1.40	3.50	2.10	3.40	2.00	3.00	1.60	3.50	2.10	3.40	2.00
13	1.40	2.30	0.90	2.20	0.80	1.70	0.30	2.30	0.90	2.20	0.80
14	2.10	1.50	−0.60	1.40	−0.70	0.80	−1.30	1.40	−0.70	1.40	−0.70
17	0.70	0.30	−0.40	0.30	−0.40	0.10	−0.60	0.30	−0.40	0.30	−0.40
RSDS		7.56		7.73		7.82		7.08		7.16	

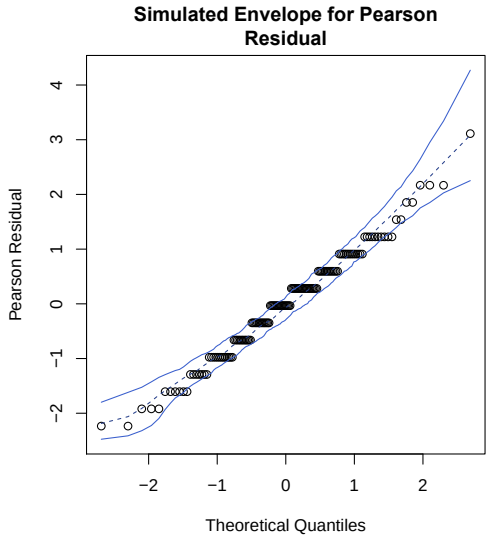
(a) NB model



(b) PIG model



(c) ZINB model



(d) ZIPIG model

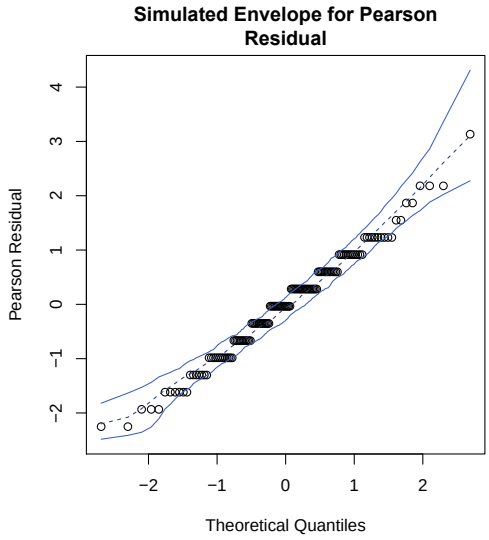


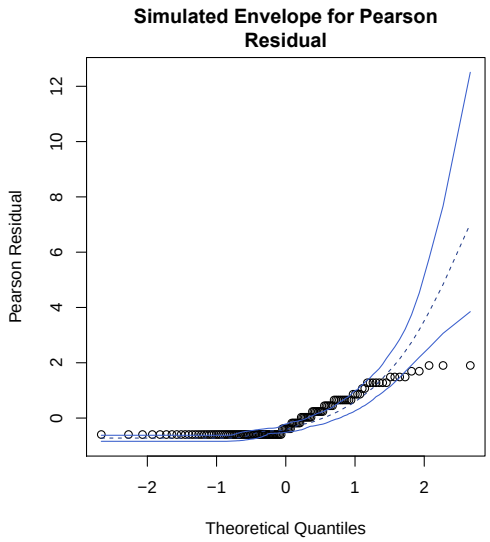
Figure 11: Simulated envelopes for the Pearson residual in the regression models under 8 hours photoperiod

However, when we observe the second one, the data set under the 16 hours photoperiod, it is possible to conclude that the ZINB and the ZIPIG models are more adequate, while the NB and the PIG models are not suitable. Those conclusions are reflexes from table 22, wherein the smaller RSDS comes from the ZINB and the ZIPIG regression models and we can see this also at figure 12, wherein the residuals remain almost all inside the envelopes only for the ZINB and the ZIPIG regression models, while for the NB and the PIG regression models one may notice many residuals out of the envelopes.

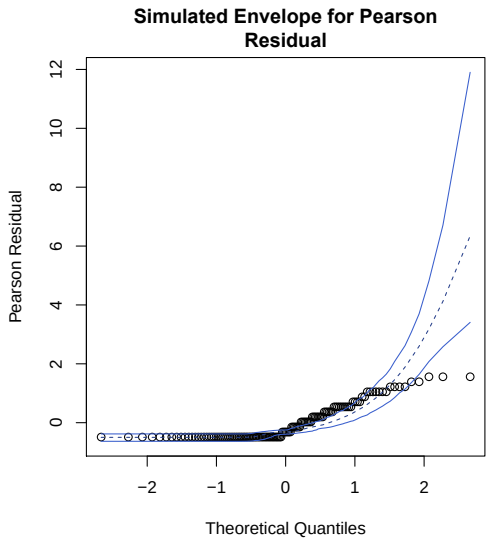
Table 22: Observed x Predicted values of apple roots count under 16 hours photoperiod

16 hrs photoperiod		NB		PIG		ZIP		ZINB		ZIPIG	
w	Observed	Predic	Diff	Predic	Diff	Predic	Diff	Predic	Diff	Predic	Diff
0	47.70	43.00	−4.70	37.30	−10.40	47.70	0.00	47.70	0.00	47.70	0.00
1	5.40	15.20	9.80	22.20	16.80	1.20	−4.20	3.90	−1.50	3.50	−1.90
2	5.40	9.40	4.00	11.90	6.50	3.40	−2.00	5.90	0.50	5.80	0.40
3	6.20	6.60	0.40	7.00	0.80	6.10	−0.10	6.90	0.70	7.10	0.90
4	6.20	4.90	−1.30	4.60	−1.60	8.30	2.10	7.00	0.80	7.30	1.10
5	4.60	3.80	−0.80	3.20	−1.40	9.00	4.40	6.40	1.80	6.70	2.10
6	7.70	3.00	−4.70	2.30	−5.40	8.20	0.50	5.50	−2.20	5.60	−2.10
7	3.10	2.40	−0.70	1.80	−1.30	6.40	3.30	4.50	1.40	4.50	1.40
8	1.50	1.90	0.40	1.40	−0.10	4.30	2.80	3.50	2.00	3.40	1.90
9	5.40	1.60	−3.80	1.10	−4.30	2.60	−2.80	2.60	−2.80	2.50	−2.90
10	3.10	1.30	−1.80	0.90	−2.20	1.40	−1.70	1.90	−1.20	1.80	−1.30
11	1.50	1.10	−0.40	0.80	−0.70	0.70	−0.80	1.40	−0.10	1.30	−0.20
12	2.30	0.90	−1.40	0.60	−1.70	0.30	−2.00	0.90	−1.40	0.90	−1.40
RSDS		13.39		22.26		8.93		5.38		5.68	

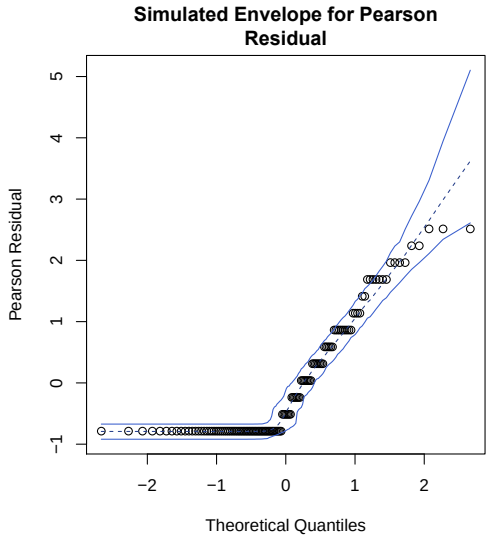
(a) NB model



(b) PIG model



(c) ZINB model



(d) ZIPIG model

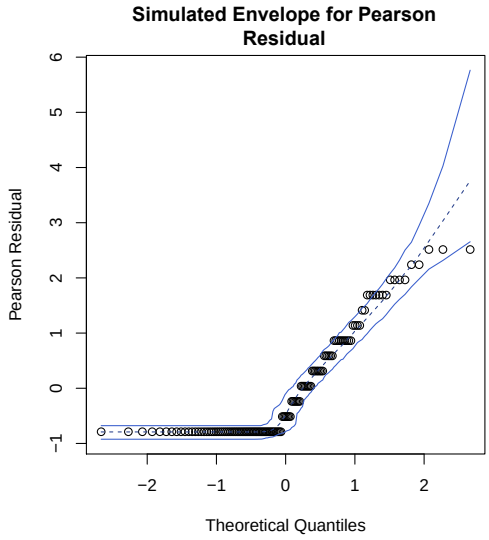


Figure 12: Simulated envelopes for the Pearson residual in the regression models under 16 hours photoperiod

6 Conclusion

In this master's thesis work a general class of zero-inflated mixed Poisson regression models was proposed based on a mixing between the general mixed Poisson distributions (that arise from a mixing between the Poisson distribution and a distribution belonging to the continuous exponential family) and a point mass of one at zero. Thereby, distinct zero-inflated overdispersed models, which have been studied in an isolated manner were unified. Furthermore, the proposed class opened the possibility of new models arise.

After all work previously developed, it is possible to conclude that the main goal was reached. In other words, based on the results presented, the proposed model can deal with overdispersion and the excess of zero in count data, since satisfactory results were reached even for samples of size 50 in scenarios with several parameters. Furthermore, we draw attention to the importance of the proposed model and its techniques to obtain the model estimates, once one could observe good results from the EM algorithm estimates, while GAMLSS models, in several cases, could not be fitted or produced unpleasant estimates.

We also pay attention that when the practitioner is dealing with count data, he must be attentive for overdispersion and zeros excess, otherwise all results and conclusions can be compromised, as we could notice through the empirical illustration, wherein we could realize the importance of the diagnostics analysis. Therefore, stem from that analysis, we may say that the NB and the PIG regression models are suitable for dealing with overdispersed count data sets that do not have zeros excess, however the ZIMP regression models have been fitted just well as those models, at least for the used data set, producing γ's parameters that conduct the zero-inflation parameter τ near to zero.

Besides of this work, an important contribution for the statistical researches is the computational tools development. With this in mind, we believe that a package that summarizes the work here developed, as well as provides appropriated support for dealing with zero-inflated and overdispersed count data, should be designed. For this reason, we hope develop a R software package soon.

Appendix A

This section begins with corrections of some typing errors in Barreto-Souza and Simas (2016). In that paper, in the second section (The model, p. 3), the exponent of the term $[\phi(\phi+2\mu)]^{-(y-1/2)}$ should be viewed as $[\phi(\phi+2\mu)]^{-(y-1/2)/2}$.

Also in that paper, in the third section (EM algorithm, p. 4), one can find the expression, in **Proposition 1**, for the conditional expectation $E(g(Z)|Y=y)$, wherein is necessary to add the $b(\xi_0)$ term. Thus, the correct expression is

$$E(g(Z)|Y=y) = \frac{d\ p(y;\ \mu_t^*,\ \phi + t/dt|_{t=0})}{p(y;\ \mu,\ \phi)} - d'(\phi) - \xi_0 + b(\xi_0)$$

instead of

$$E(g(Z)|Y=y) = \frac{d\ p(y;\ \mu_t^*,\ \phi + t/dt|_{t=0})}{p(y;\ \mu,\ \phi)} - d'(\phi) - \xi_0.$$

Moreover, in the **Proof of Proposition 1** (Appendix, p. 16) there is the expression for the conditional moment generating function of $g(Z)$ and where one reads $n!$ it should be read as $y!$. In addition, the term ϕ inside of the exponential and before the term $b\left(\frac{\phi\xi_0}{\phi+t}\right)$ should be replaced by $(\phi+t)$. Thus, the correct expression is

$$\begin{aligned} E(\exp\{tg(Z)\}|Y=y) &= \frac{\mu^y}{y!p(y;\mu,\phi)}\int_0^\infty e^{-\mu z}z^y \times \exp\{\phi[z\xi_0 - b(\xi_0)] \\ &\quad + d(\phi) + (\phi+t)g(z) + h(z)\}dz \\ &= \exp\left\{(\phi+t)b\left(\frac{\phi\xi_0}{\phi+t}\right) - d(\phi+t) + d(\phi) - \phi b(\xi_0)\right\} \\ &\quad \times \frac{p(y;\mu_t^*,\phi+t)}{p(y;\mu,\phi)}, \end{aligned}$$

instead of

$$\begin{aligned} E(\exp\{tg(Z)\}|Y=y) &= \frac{\mu^y}{n!p(y;\mu,\phi)}\int_0^\infty e^{-\mu z}z^y \times \exp\{\phi[z\xi_0 - b(\xi_0)] \\ &\quad +d(\phi)+(\phi+t)g(z)+h(z)\}dz \\ &= \exp\left\{\phi b\left(\frac{\phi\xi_0}{\phi+t}\right) - d(\phi+t)+d(\phi)-\phi b(\xi_0)\right\} \\ &\quad \times\frac{p(y;\mu_t^*,\phi+t)}{p(y;\mu,\phi)}. \end{aligned}$$

Another typing error that should be corrected, also in the Appendix, p. 16, remains in the **Proof of Proposition 2**. Where one finds z_{il}, in the expression for the expectation $E\left(\frac{\partial l_c}{\partial\beta_j}\frac{\partial l_c}{\partial\alpha_l}|Y\right)$, the correct term is w_{il}.

In what follows, we present the proof of proposition 1 given in the EM algorithm chapter.

Proof of Proposition 1

We have that

$$\begin{aligned} E(B \mid W=w) &= \sum_{b=0}^{1}\frac{b\ P(W=w \mid B=b)\ P(B=b)}{P(W=w)} \\ &= \frac{P(W=w \mid B=1)\ P(B=1)}{P(W=w)} \\ &= \frac{(1-\tau)\ p_Y(w;\ \mu,\ \phi)}{p_W(w;\ \mu,\ \phi,\ \tau)}, \end{aligned}$$

$$
\begin{aligned}
E(BZ \mid W = w) &= \int_0^\infty \sum_{b=0}^{1} \frac{bz\ P(W = w \mid Z = z, B = b)\ f_Z(z)\ P(B = b)}{P(W = w)}\ dz \\
&= \int_0^\infty \frac{z\ P(W = w \mid Z = z, B = 1)\ f_Z(z)\ P(B = 1)}{P(W = w)}\ dz \\
&= \int_0^\infty \frac{z\ P(Y = w \mid Z = z)\ f_Z(z)\ (1 - \tau)}{P(W = w)}\ dz \\
&= \frac{(1 - \tau)\ p_Y(w;\ \mu,\ \phi)}{p_W(w;\ \mu,\ \phi,\ \tau)} \int_0^\infty z\ f(Z = z \mid Y = y)\ dz \\
&= \frac{(1 - \tau)\ p_Y(w;\ \mu,\ \phi)}{p_W(w;\ \mu,\ \phi,\ \tau)}\ E(Z \mid Y),
\end{aligned}
$$

and

$$
\begin{aligned}
E(Bg(Z) \mid W = w) &= \int_0^\infty \sum_{b=0}^{1} \frac{bg(z)\ P(W = w \mid Z = z, B = b)\ f_Z(z)\ P(B = b)}{P(W = w)}\ dz \\
&= \int_0^\infty \frac{g(z)\ P(W = w \mid Z = z, B = 1)\ f_Z(z)\ P(B = 1)}{P(W = w)}\ dz \\
&= \int_0^\infty \frac{g(z)\ P(Y = w \mid Z = z)\ f_Z(z)\ (1 - \tau)}{P(W = w)}\ dz \\
&= \frac{(1 - \tau)\ p_Y(w;\ \mu,\ \phi)}{p_W(w;\ \mu,\ \phi,\ \tau)} \int_0^\infty g(z)\ f(Z = z \mid Y = y)\ dz \\
&= \frac{(1 - \tau)\ p_Y(w;\ \mu,\ \phi)}{p_W(w;\ \mu,\ \phi,\ \tau)}\ E(g(Z) \mid Y).
\end{aligned}
$$

Proposition 2 *Let $W \sim ZIMP(\mu, \phi, \tau)$, with $Z \sim EF(\xi_0, \phi)$ and $B \sim Bernoulli(\tau)$, the previous latent variables defined, and $Y \sim MP(\mu, \phi)$. Thus,*

$$E(B^2|W) = (1-\tau)\frac{p_Y(w;\mu,\phi)}{p_W(w;\mu,\phi,\tau)},$$

$$E(B^2Z|W) = \frac{(1-\tau)(w+1)}{\mu}\frac{p_Y(w+1;\mu,\phi)}{p_W(w;\mu,\phi,\tau)},$$

$$E(B^2Z^2|W) = \frac{(1-\tau)(w+1)(w+2)}{\mu^2}\frac{p_Y(w+2;\mu,\phi)}{p_W(w;\mu,\phi,\tau)},$$

$$E(B^2g(Z)|W) = (1-\tau)\frac{p_Y(w;\mu,\phi)}{p_W(w;\mu,\phi,\tau)}\left(\frac{dp_Y(w;\mu_t^*,\phi+t)/dt|_{t=0}}{p_Y(w;\mu,\phi)} - d'(\phi) - \xi_0 + b(\xi_0)\right),$$

$$\begin{aligned} E(B^2g^2(Z)|W) &= (1-\tau)\frac{p_Y(w;\mu,\phi)}{p_W(w;\mu,\phi,\tau)}\left\{[d'(\phi)+\xi_0]^2 - 2[d'(\phi)+\xi_0]b(\xi_0) + \frac{\xi_0^2 b''(\xi_0)}{\phi}\right. \\ &+ 2[b(\xi_0) - d'(\phi) - \xi_0]\frac{dp_Y(w;\mu_t^*,\phi+t)/dt|_{t=0}}{p_Y(w;\mu,\phi)}b^2(\xi_0) - d''(\phi) \\ &+ \left.\frac{d^2p_Y(w;\mu_t^*,\phi+t)/dt^2|_{t=0}}{p_Y(w;\mu,\phi)}\right\}, \end{aligned}$$

$$E(B^2Zg(Z)|W) = \frac{(1-\tau)(w+1)}{\mu}\frac{p_Y(w+1;\mu,\phi)}{p_W(w;\mu,\phi,\tau)}\left(\frac{dp_Y(w+1;\mu_t^*,\phi+t)/dt|_{t=0}}{p_Y(w+1;\mu,\phi)} - d'(\phi) - \xi_0 + b(\xi_0)\right).$$

Appendix B

Now, we present in table 23 the time, in minutes, of EM algorithm to produces the estimates of the 4500 samples runs of the Monte Carlo study for the ZIMP regression models, presented in the simulation study chapter, performed in a 64-bit version of Windows 7, an Intel Core i7 @ 3.4 GHz and 8GB RAM.

Table 23: EM algorithm time in minutes

ZINB	10% zero-inflation	30% zero-inflation	50% zero-inflation
$n = 50$	86	71	102
$n = 100$	81	88	129
$n = 200$	137	148	171
$n = 300$	200	216	222
ZIPIG	10% zero-inflation	30% zero-inflation	50% zero-inflation
$n = 50$	242	240	295
$n = 100$	213	239	246
$n = 200$	323	334	307
$n = 300$	445	452	381

References

Barreto-Souza, W. and Simas, A. B. (2016). General mixed poisson regression models with varying dispersion, *Statistics and Computing* **26**(6): 1263–1280.

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L. and Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162**(2): 195–209.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*, Cambridge University Press.

Cook, R. D. (1977). Detection of influential observation in linear regression, *Technometrics* **19**(1): 15–18.

Dean, C. B. and Nielsen, J. D. (2007). Generalized linear mixed models: a review and some extensions, *Lifetime Data Analysis* **13**(4): 497–512.

Dean, C., Lawless, J. F. and Willmot, G. E. (1989). A mixed poisson-inverse gaussian regression model, *Canadian Journal of Statistics* **17**(2): 171–181.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1): 1–38.

Famoye, F. and Singh, K. P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data, *Journal of Data Science* **4**(1): 117–130.

Garay, A. M., Hashimoto, E. M., Ortega, E. M. and Lachos, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models, *Computational Statistics & Data Analysis* **55**(3): 1304 – 1318.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study, *Biometrics* **56**(4): 1030–1039.

Hilbe, J. M. (2014). *Modeling Count Data*, Cambridge University Press, Cambridge.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: Models and estimation, *Computational Statistics & Data Analysis* **27**(2): 151 – 170.

Holla, M. S. (1967). On a poisson-inverse gaussian distribution, *Metrika* **11**(1): 115–121.

Karlis, D. (2001). A general em approach for maximum likelihood estimation in mixed poisson regression models, *Statistical Modelling* **1**(4): 305–318.

Karlis, D. and Xekalaki, E. (2005). Mixed poisson distributions, *International Statistical Review* **73**(1): 35–58.

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

Lawless, J. F. (1987). Negative binomial and mixed poisson regression, *Canadian Journal of Statistics* **15**(3): 209–225.

Lee, A. H., Wang, K. and Yau, K. K. (2001). Analysis of zero-inflated poisson data incorporating extent of exposure, *Biometrical Journal* **43**(8): 963–975.

Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P. A. and Peterson, J. P. (1999). Multivariate zero-inflated poisson models and their applications, *Technometrics* **41**(1): 29–38.

Lim, H. K., Li, W. K. and Yu, P. L. (2014). Zero-inflated poisson regression mixture model, *Computational Statistics & Data Analysis* **71**: 151 – 158.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **44**(2): 226–233.

Mwalili, S. M., Lesaffre, E. and Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research, *Statistical Methods in Medical Research* **17**(2): 123–139. PMID: 17698937.

Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. and Rothkamm, K. (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study, *Biometrical Journal* **58**(2): 259–279.

Ridout, M., Demétrio, C. G. and Hinde, J. (1998). Models for count data with many zeros, *Proceedings of the XIXth International Biometrics Conference, Cape Town , Invited Papers* **19**: 179–192.

Ridout, M., Hinde, J. and Demétrio, C. G. B. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives, *Biometrics* **57**(1): 219–223.

Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data, *The Annals of Applied Statistics* **4**(2): 943–961.

Shankar, V., Milton, J. and Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry, *Accident Analysis & Prevention* **29**(6): 829 – 837.

Stasinopoulos, D. and Rigby, R. (2007). Generalized additive models for location scale and shape (gamlss) in r, *Journal of Statistical Software* **23**(1): 1–46.

Stoyanov, J. and Lin, G. D. (2011). Mixtures of power series distributions: identifiability via uniqueness in problems of moments, *Annals of the Institute of Statistical*

Mathematics **63**(2): 291–303.

Willmot, G. E. (1987). The poisson-inverse gaussian distribution as an alternative to the negative binomial, *Scandinavian Actuarial Journal* **1987**(3-4): 113–127.

Yau, K. K. W., Wang, K. and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, *Biometrical Journal* **45**(4): 437–452.

Zhu, H., Lee, S.-Y., Wei, B.-C. and Zhou, J. (2001). Case-deletion measures for models with incomplete data, *Biometrika* **88**(3): 727.