University of São Paulo
"Luiz de Queiroz" College of Agriculture

Contributions to the analysis of dispersed count data

Eduardo Elias Ribeiro Junior

Dissertation presented to obtain the degree of Master in Science
Area: Statistics and Agricultural Experimentation
Piracicaba, 2018

University of São Paulo
“Luiz de Queiroz” College of Agriculture

Contributions to the analysis of dispersed count data

Eduardo Elias Ribeiro Junior

Dissertation presented to obtain the degree of Master in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2019

Eduardo Elias Ribeiro Junior
Bachelor in Statistics

Contributions to the analysis of dispersed count data

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Profa. Dra. **CLARICE GARCIA BORGES DEMÉTRIO**

Dissertation presented to obtain the degree of Master in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Ribeiro Junior, Eduardo Elias
Contributions to the analysis of dispersed count data / Eduardo Elias Ribeiro Junior. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2019 .
82 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Dados de contagens 2. Dispersão variável 3. Inferência baseada em verossimilhança 4. Modelos probabilísticos discretos 5. Subdispersão 6. Superdipersão . I. Título.

To my family

CONTENTS

RESUMO

Contribuições à análise de dados de contagem

Em diversos estudos agrícolas e biológicos, a variável resposta é um número inteiro não negativo que desejamos explicar ou analisar em termos de um conjunto de covariáveis. Diferentemente do modelo linear Gaussiano, a variável resposta é discreta com distribuição de probabilidade definida apenas em valores do conjunto dos naturais. O modelo Poisson é o modelo padrão para dados em forma de contagens. No entanto, as suposições desse modelo forçam que a média seja igual a variância, o que pode ser implausível em muitas aplicações. Motivado por conjuntos de dados experimentais, este trabalho teve como objetivo desenvolver métodos mais realistas para a análise de contagens. Foi proposta uma nova reparametrização da distribuição COM-Poisson e explorados modelos de regressão baseados nessa distribuição. Uma extensão desse modelo para permitir que a dispersão, assim como a média, dependa de covariáveis, foi proposta. Um conjunto de modelos para contagens, nomeadamente COM-Poisson, *Gamma-count*, Weibull discreto, Poisson generalizado, duplo Poisson e Poisson-Tweedie, foi revisado e comparado, considerando os índices de dispersão, inflação de zero e cauda pesada, juntamente com os resultados de análises de dados. As rotinas computacionais desenvolvidas nesta dissertação foram organizadas em dois pacotes `R` disponíveis no GitHub.

Keywords: Dados de contagens, Dispersão variável, Inferência baseada em verossimilhança, Modelos probabilísticos discretos, Subdispersão, Superdipersão.

ABSTRACT

Contributions to the analysis of dispersed count data

In many agricultural and biological contexts, the response variable is a nonnegative integer value which we wish to explain or analyze in terms of a set of covariates. Unlike the Gaussian linear model, the response variable is discrete with a distribution that places probability mass at natural numbers only. The Poisson regression is the standard model for count data. However, assumptions of this model forces the equality between mean and variance, which may be implausible in many applications. Motivated by experimental data sets, this work intended to develop more realistic methods for the analysis of count data. We proposed a novel parametrization of the COM-Poisson distribution and explored the regression models based on it. We extended the model to allow the dispersion, as well as the mean, depending on covariates. A set of count statistical models, namely COM-Poisson, Gamma-count, discrete Weibull, generalized Poisson, double Poisson and Poisson-Tweedie, was reviewed and compared, considering the dispersion, zero-inflation, and heavy tail indexes, together with the results of data analyzes. The computational routines developed in this dissertation were organized in two `R` packages available on GitHub.

Keywords: Count data, Discrete probability models, Likelihood-based inference, Overdispersion, , Underdispersion, Varying dispersion.

1 GENERAL INTRODUCTION

Count data arise from random variables that take non-negative integer values and typically represent the number of times an event occurs in an observation period or region. This kind of data is also common in crop sciences, examples including the number of grains produced by a plant, the number of fruits on a tree, the number of insects captured by a trap, etc. Since the seminal paper of Nelder and Wedderburn (1972), where the class of the generalized linear models (GLMs) was introduced, a Poisson regression model has been often used for the analysis of count data. This model provides a suitable strategy for the analysis of count data and an efficient Fisher scoring algorithm that can be used for fitting.

In spite of the advantages of the Poisson regression model, the Poisson distribution has only one parameter that represents both the expectation and variance of the count random variable. This restriction on the relationship between the expectation and variance of the Poisson distribution is referred as equidispersion. However, in practical data analysis equidispersion can be unsuitable, since the observed data can present variance both smaller or larger than the mean, leading to under- and overdispersion, respectively. In both cases, the Poisson model estimates the regression coefficients consistently, but their associated standard errors are inconsistent, which in turn can lead to misleading inferences (Winkelmann and Zimmermann 1994; Bonat et al. 2018).

In practice, overdispersion is widely reported in the literature and may occur due to the absence of relevant covariates, heterogeneity of sampling units, different observational periods/regions not being considered in the analysis, and excess of zeros (Hinde and Demétrio 1998). Underdispersion is less often reported in the literature, however, it has been of increasing interest in the statistical community. The processes that reduce the variability are not as well-known as those leading to extra variability. For this reason, there are few models to deal with underdispersed count data. Possible explanatory mechanisms leading to underdispersion may be related to the underlying stochastic process generating the count data. For example, when the time between events is not exponentially distributed, the number of events can be over- or underdispersed; a process that motivated the duration dependence in count data models (Winkelmann 1995). Another possible explanation for underdispersion is when the responses correspond to order statistics of component observations, such as maxima of Poisson distributed counts (Steutel and Thiemann 1989).

Strategies for constructing alternative count distributions are related to the causes of the non-equidispersion. Specifically for overdispersion, Poisson mixture (compound) models are widely applied. One popular example of this approach is the negative-binomial model, where the expectation of the Poisson distribution is assumed to be gamma distributed (Hinde and Demétrio 1998). However, other distributions can also be used to represent this additional random variation. For example the Poisson-Tweedie model (Bonat et al. 2018) and its special cases Poisson inverse-Gaussian and Neyman-Type A models assume that the random effects are Tweedie, inverse Gaussian or Poisson distributed, respectively. The Gamma-count distribution assumes a gamma distribution for the time between events and it can handle underdispersed as well as overdispersed count data (Zeviani et al. 2014). The COM-Poisson distribution, is obtained by a generalization of the Poisson distribution that allows for a non-linear decrease in the ratios of successive probabilities (Shmueli et al. 2005) and can be seen as a particular case of the weigthed Poisson distribution (Del Castillo and Pérez-Casany 1998). The discretization process can also be used to derive flexible discrete distributions, such as the discrete Weibull distribution (Nakagawa and Osaki 1975; Klakattawi, Vinciotti, and Yu 2018). Related to negative binomial distribution, the generalized Poisson is obtained as a limiting form of generalized negative binomial distribution (Consul and Jain 1973). The double exponential family and the particular double Poisson (Efron 1986) can model equi-, under- and overdispersion in count data, as well. A comprehensive discussion on other generalizations of the Poisson distribution can be found in Winkelmann (2008).

The standard regression for the models from the exponential family (generalized linear models) and for the aforementioned models is linked to the mean (or location-related) parameter. Thereby, the variance of the response variable is completely specified by the variance function (be it known or not). However, this assumption cannot be reasonable. For example, when there is an effect of a covariate (treatment, dose, etc.) in both expectation and dispersion of the count random variable. Smyth (1988) shows how to include a linear predictor for the dispersion as well as for the mean in the generalized linear models. In his paper, Smyth illustrates this methodology with continuous data. For discrete data, his proposal can be extended by using quasi-likelihood estimation methods (Wedderburn 1974; Nelder and Pregibon 1987). Besides that, a full parametric approach is presented by Rigby and Stasinopoulos (2005) using the so-called generalized additive models for location, scale, and shape (GAMLSS).

The main objective of this dissertation consists of exploring novel modeling approaches for the analysis of count data. The remainder of the text is organized as follow. In Chapter 2, the motivating datasets, mostly from biological experiments, and its challenges for analysis are presented. Chapter 3 is devoted to present and study our novel reparametrization of COM-Poisson models. An overview and comparison of several flexible probability distributions for count data are addressed in Chapter 4. In Chapter 5, we consider varying dispersion in the reparametrized COM-Poisson models where the mean and the dispersion parameters may be allowed to depend on covariates. Final considerations are given in Chapter 6.

References

Bonat, W. H., B. Jørgensen, C. C. Kokonendji, and J. Hinde (2018). "Extended Poisson-Tweedie: properties and regression model for count data". In: *Statistical Modelling* 18.1, pp. 24–49.

Consul, P. C. and G. C. Jain (1973). "A Generalization of the Poisson Distribution". In: *Technometrics* 15.4, pp. 791–799.

Del Castillo, J. and M. Pérez-Casany (1998). "Weighted Poisson Distributions for Overdispersion and Underdispersion Situations". In: *Annals of the Institute of Statistical Mathematics* 50.3, pp. 567–585.

Efron, B. (1986). "Double Exponential Families and Their Use in Generalized Linear Regression". In: *Journal of the American Statistical Association* 84.395, pp. 709–721.

Hinde, J. and C. G. B. Demétrio (1998). "Overdispersion: models and estimation". In: *Computational Statistics & Data Analysis* 27.2, pp. 151–170.

Klakattawi, H. S., V. Vinciotti, and K. Yu (2018). “A Simple and Adaptive Dispersion Regression Model for Count Data”. In: *Entropy* 20.142.

Nakagawa, T. and S. Osaki (1975). “The Discrete Weibull Distribution”. In: *IEEE Transactions on Reliability* 24.5, pp. 300–301.

Nelder, J. A. and D. Pregibon (1987). “An Extended Quasi-likelihood Function”. In: *Biometrika* 74.2, pp. 221–232.

Nelder, J. A. and R. W. M. Wedderburn (1972). “Generalized Linear Models”. In: *Journal of the Royal Statistical Society. Series A (General)* 135, pp. 370–384.

Rigby, R. A. and D. M. Stasinopoulos (2005). “Generalized additive models for location, scale and shape (with discussion)”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.3, pp. 507–554.

Shmueli, G., T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright (2005). “A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.1, pp. 127–142.

Smyth, G. K. (1988). “Generalized Linear Models with Varying Dispersion”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 51.1, pp. 47–60.

Steutel, F. W. and J. G. F. Thiemann (1989). “The gamma process and the Poisson distribution”. In: *(Memorandum COSOR; Vol. 8924). Eindhoven: Technische Universiteit Eindhoven.*

Wedderburn, R. W. M. (1974). “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method”. In: *Biometrika* 61.3, p. 439.

Winkelmann, R. (1995). “Duration Dependence and Dispersion in Count-Data Models”. In: *Journal of Business & Economic Statistics* 13.4, pp. 467–474.

Winkelmann, R. (2008). *Econometric Analysis of Count Data.* 5th edition. Berlin, Heidelberg: Springer-Velag, p. 342.

Winkelmann, R. and K. F. Zimmermann (1994). “Count Data Models for Demographic Data”. In: *Mathematical Population Studies* 4.3, pp. 205–221.

Zeviani, W. M., P. J. Ribeiro Jr, W. H. Bonat, S. E. Shimakura, and J. A. Muniz (2014). “The Gamma-count distribution in the analysis of experimental underdispersed data”. In: *Journal of Applied Statistics* 41.12, pp. 2616–2626.

2 MOTIVATING STUDIES

In this chapter, we present a number of data sets to illustrate some scientific and statistical issues which arise from count data. These data sets will be used throughout the text for illustration of the proposed methodologies. All data sets are available in the packages `cmpreg` and `flexcm` for the software `R` (R Core Team 2018).

2.1 Artificial defoliation in cotton phenology

Cotton production can be drastically reduced by attack of defoliating insects. Depending on the growth stage, the plants can recover from the caused damage and keep production not affected or can have the production reduced by low intensity defoliation. In order to study the recovery of cotton plants (*Gossypium hirsutum*) in terms of production, Silva et al. (2012) conducted a greenhouse experiment under a completely randomized design with five replicates. The experimental unity was a pot with two plants and it was recorded the number of cotton bolls at five artificial defoliation levels (0%, 25%, 50%, 75%, and 100%) and five growth stages: vegetative, flower-bud, blossom, boll and boll open.

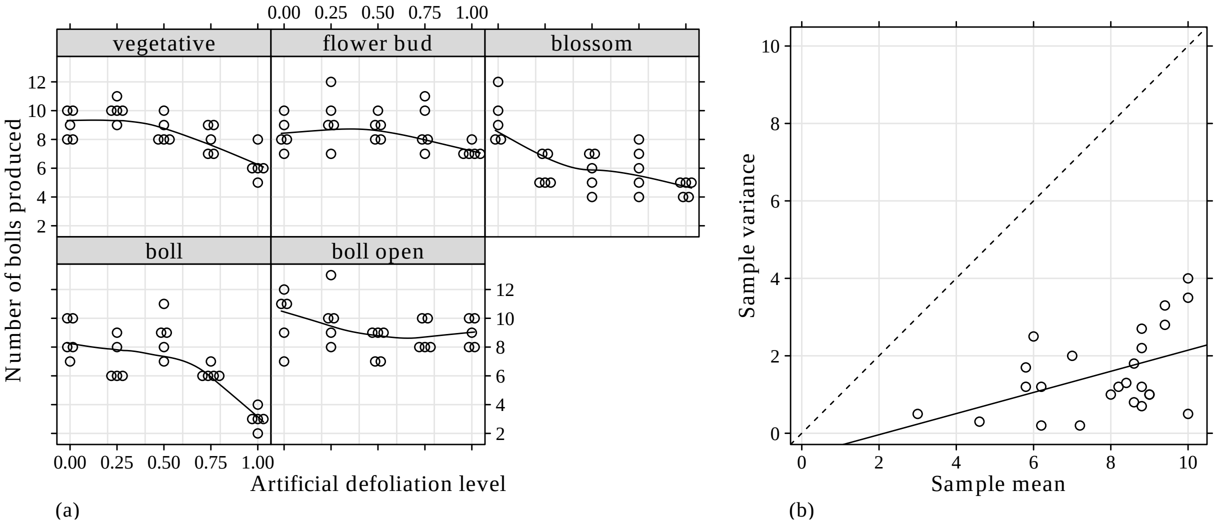

Figure 2.1. (a) Number of bolls produced for each artificial defoliation level and each growth stage (solid lines represent loess curves) and (b) sample variance against the sample mean of the five replicates for each combination of defoliation level and growth stage (dotted line is the identity line and solid line is the least square line).

Figure 2.1(a) shows the number of cotton bolls recorded for each combination of defoliation level and growth stage; the smoothers indicate that are different quadratic effects by growth stage. Figure 2.1(b) show that all sample variances are smaller than the sample means, suggesting underdispersion.

Alternatives analysis of this data have been proposed in the literature. Zeviani et al. (2014) analyzed it using the Gamma-count distribution. Bonat et al. (2018) used this data for illustrating the extended Poisson-Tweedie model. Huang (2017) and Ribeiro Jr et al. (2018) analyzed it by using different mean-parametrizations of the COM-Poisson model.

2.2 Soil moisture and potassium fertilization on soybean culture

In this second example we consider a study of potassium doses and soil moisture levels on soybean (*Glicine Max*) production. The tropical soils are usually poor in potassium (K) and demand potassium fertilization to obtain satisfactory yields when cultivated soybean. Furthermore, soybean production is affected by long exposition to water deficit. As potassium is a nutrient involved in the water balance in plant, by hyphotesis, a good supply of potassium avoids to reduce production.

To evaluate the effects of potassium doses and soil humidity levels on soybean production, Serafim et al. (2012) conducted a 5 × 3 factorial experiment in a randomized complete block design with 5 replicates. Five different potassium doses (0, 0.3, 0.6, 1.2 and 1.8 × 100mg dm^{-3}) were applied to the soil and soil moisture levels were controlled at (37.5, 50, and 62.5%). The experiment was carried out in a greenhouse and the experimental units were pots with two plants in each. The count responses measured were the total number of bean seeds per pot and the number of pods.

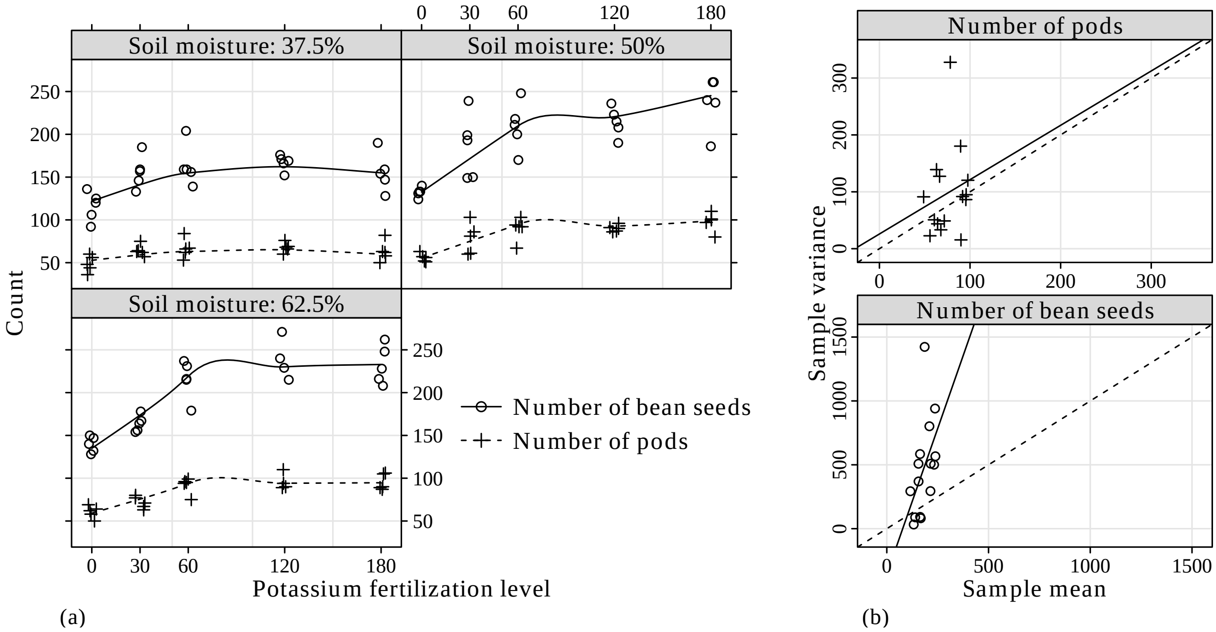

Figure 2.2. (a) Number of bean seeds and number of pods produced for each moisture level and each potassium dose (solid lines represent loess curves) and (b) sample variance against the sample mean of the five replicates for each combination of moisture level and potassium dose (dotted line is the identity line and solid line is the least-squares line).

Figure 2.2(a) shows the number of bean seeds and the number of pods recorded for each combination of potassium dose and moisture level; it is important to note the indication of a quadratic effect of the potassium levels for both counts, as indicated by the smoothers. For the number of bean seeds, most points in Figure 2.2(b) are above the identity line, suggesting overdispersion (block effect not yet removed). However, for the number of pods, there are points above and below the identity line – that leads to least-squares line very similar to the identity line, suggesting that the equidispersion assumption can be suitable, if the experimental conditions are not related to the dispersion.

2.3 Toxicity of nitrofen in aquatic systems

Nitrofen is a herbicide that was used extensively for the control of broad-leaved and grass weeds in cereals and rice. Although it is relatively non-toxic to adult mammals, nitrofen is a significant teratogen and mutagen. It is also acutely toxic and reproductively toxic to cladoceran zooplankton. Nitrofen is no longer in commercial use in the United States, having been the first pesticide to be withdrawn due to tetragenic effects (Bailer and Oris 1994).

This data set comes from an experiment to measure the reproductive toxicity of the herbicide, nitrofen, on a species of zooplankton (*Ceriodaphnia dubia*). Fifty animals were randomized into batches of ten and each batch was put in a solution with a measured concentration of nitrofen ($0, 0.8, 1.6, 2.35$ and 3.10 μg/100litre) (`dose`). Subsequently, the number of live offspring was recorded.

Table 2.1. Sample means, sample variances and sample dispersion indexes of the ten replicates for each concentration level in the nitrofen study ($\text{DI} = \bar{x}/s^2$).

Dose	N. obs.	Sample mean	Sample variance	Sample DI
0.00	10	32.40	13.1556	0.4060
0.80	10	31.50	10.7222	0.3404
1.60	10	28.30	5.5667	0.1967
2.35	10	17.20	34.8444	2.0258
3.10	10	6.00	13.7778	2.2963

Table 2.1 shows the sample mean, sample variance and sample dispersion index of the number of live offspring obtained from the batches with different nitrofen concentration level. This descriptive analysis indicates the nitrofen reduce the number of live offsprings however it seems like the dispersion is also influenced by the nitrofen concentration level (doses up to 1.6μg/10^2litre suggesting underdispersion while doses between 1.6 and 3.1μg/10^2litre suggesting overdispersion). The Figure 2.3 shows a graphical representation of these results.

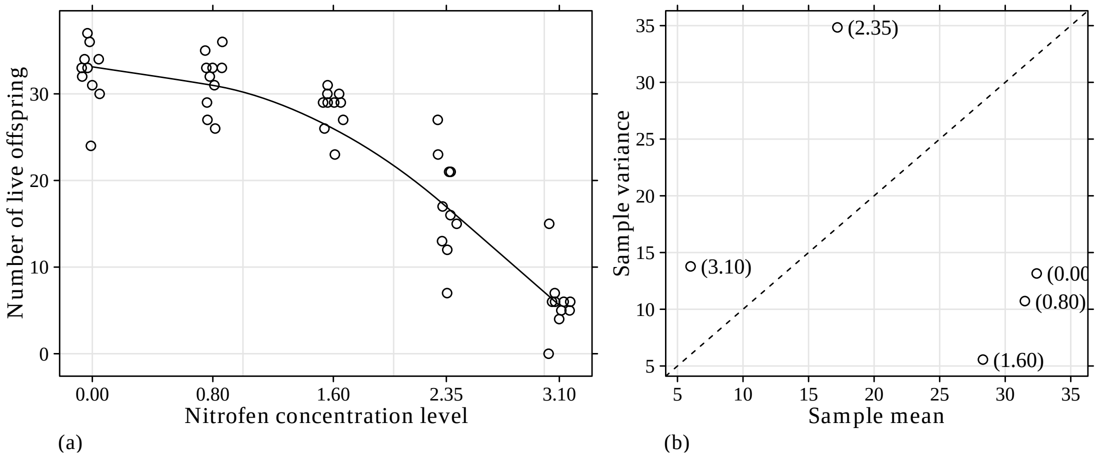

Figure 2.3. (a) Number of live offsprings observed for each nitrofen concentration level and (b) scatterplot of the sample means against sample variances (dotted line is the identity line and solid line is the least-squares line).

2.4 *Annona mucosa* in control of stored maize peast

New control methods are necessary for stored grain pest management programs due to both the widespread problems of insecticide-resistance populations and the increasing concerns of consumers regarding pesticide residues in food products. Ribeiro et al. (2013) carried out an experiment to assess the bioactivity of extracts of *Annona mucosa* (Annonaceae) for control *Sitophilus zeamaus* (Coleoptera: Curculionidae), a major pest of stored maize in Brazil. Petri dishes containing 10g of corn were treated with extracts prepared with different parts of *Annona mucosa* (seeds, leaves and branches) or just water (control) were completely randomized with 10 replicates. Then 20 *Sitophilus zeamaus* adults were placed in each Petri dish and the numbers of emerged insects (progeny) after 60 days were recorded.

Table 2.2. Sample means, sample variances and sample dispersion indexes of the ten replicates for each treatment in the maize pest study (DI $= \bar{x}/s^2$).

Treatment	N. obs.	Sample mean	Sample variance	Sample DI
Control	10	31.50	62.5000	1.9841
Leaf extract	10	31.30	94.0111	3.0035
Branch extract	10	29.90	88.7667	2.9688
Seed extract	10	1.10	1.6556	1.5051

For all treatments, the sample variance of the number of emerged insects treatment is greater than their respective sample average, a strong indication of overdispersion. For leaf extract, the sample variance is three times higher than the mean.

This data set is presented and analysed by Ribeiro et al. (2013) and later used by Demétrio, Hinde, and Moral (2014) for illustrating the quasi-Poisson approach for modelling overdispersed count data.

2.5 Alternative substrats for bromeliad production

Xaxim is a substrate used in the cultivation of bromeliads and orchids, whose commercialization was prohibited in 2001. Since then, there has been researching in botany to propose alternative substrates to Xaxim in the cultivation of bromeliads, orchids, and other epiphytes (Salvador 2008). This dataset comes from a randomized experiment conducted in a greenhouse in four blocks design with objective of evaluate five different recipients of alternative substrates for bromeliads (Kanashiro et al. 2008). All treatments contained peat and perlite and are differed in the third component: *Pinus* bark, *Eucalyptus* bark, Coxim, coconut fiber and Xaxim. The variable of interest was the number of leaves per experimental unit (pot with initially eight plants), which was registered at 4, 173, 229, 285, 341, and 435 days after planting.

The observed number of leaves per each combination of days after planting and treatments are shown in Figure 2.4(a). There is a clear nonlinear (sigmoidal) relationship between counts and days, mainly due to the low number of leaves observed at 4 days (as expected). Figure 2.4(b) shows the sample mean and sample variances per block on the logarithm scale. All sample variances are much smaller than their sample means, even on the logarithm scale, indicating a strong evidence of underdispersion.

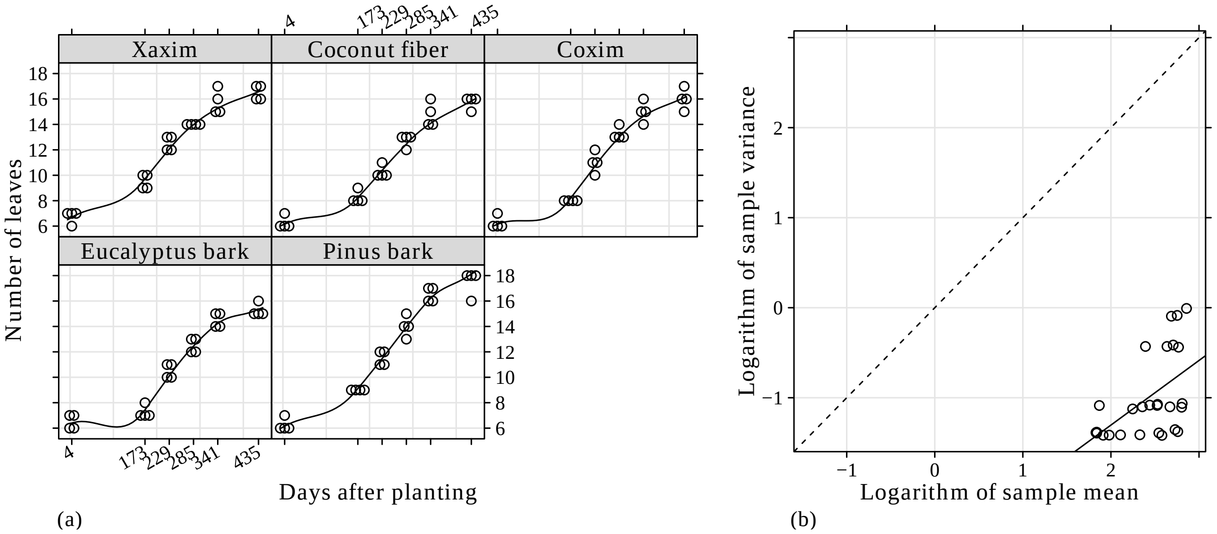

Figure 2.4. (a) Number of leaves for each treatment and day after planting (solid lines represent loess curves) and (b) sample variance against the sample mean per block (dotted line is the identity line and solid line is the least square line).

References

Bailer, A. and J. Oris (1994). “Assessing toxicity of pollutants in aquatic systems”. In: *In Case Studies in Biometry*, pp. 25–40.

Bonat, W. H., B. Jørgensen, C. C. Kokonendji, and J. Hinde (2018). “Extended Poisson-Tweedie: properties and regression model for count data”. In: *Statistical Modelling* 18.1, pp. 24–49.

Demétrio, C. G. B., J. Hinde, and R. A. Moral (2014). “Models for overdispersed data in entomology”. In: *Ecological modelling applied to entomology*. Springer, pp. 219–259.

Huang, A. (2017). “Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts”. In: *Statistical Modelling* 17.6, pp. 1–22.

Kanashiro, S., K. Minami, T. Jocys, C. T. dos Santos Dias, and A. R. Tavares (2008). “Alternative substrates to fern tree fiber in the production of ornamental bromeliad”. In: *Pesquisa Agropecuária Brasileira* 43.10.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ribeiro Jr, E. E., W. M. Zeviani, W. H. Bonat, C. G. B. Demétrio, and J. Hinde (2018). “Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data”. In: *arXiv (Statistics Applications and Statistics Methodology)*.

Ribeiro, L. P., J. D. Vendramim, K. U. Bicalho, M. S. Andrade, J. B. Fernandes, R. A. Moral, and C. G. B. Demétrio (2013). “*Annona mucosa* Jacq. (Annonaceae): A promising source of bioactive compounds against *Sitophilus zeamais* Mots. (Coleoptera: Curculionidae)”. In: *Journal of Stored Products Research* 55, pp. 6–14.

Salvador, E. (2008). “Evaluation of Alternative Substrates to the Xaxim dust for ’Matrogrossense’ fern (*Poly Aureum*) Cultivation”. In: *Acta Horticulturae* 779, pp. 547–554.

Serafim, M. E., F. B. Ono, W. M. Zeviani, J. O. Novelino, and J. V. Silva (2012). “Umidade do solo e doses de potássio na cultura da soja”. In: *Revista Ciência Agronômica* 43.2, pp. 222–227.

Silva, A. M., P. E. Degrande, R. Suekane, M. G. Fernandes, and W. M. Zeviani (2012). “Impacto de diferentes níveis de desfolha artificial nos estágios fenológicos do algodoeiro”. In: *Revista de Ciências Agrárias* 35.1, pp. 163–172.

Zeviani, W. M., P. J. Ribeiro Jr, W. H. Bonat, S. E. Shimakura, and J. A. Muniz (2014). “The Gamma-count distribution in the analysis of experimental underdispersed data”. In: *Journal of Applied Statistics* 41.12, pp. 2616–2626.

3 REPARAMETRIZATION OF COM-POISSON REGRESSION MODELS

ABSTRACT

The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution that can deal with under-, equi- and overdispersed count data. Unfortunately, its location parameter does not correspond to the expectation, which complicates the parameter interpretation. In this paper, we propose a straightforward reparametrization of the COM-Poisson distribution based on an approximation to the expectation. Estimation and inference are done using the likelihood paradigm. Simulation studies show that the maximum likelihood estimators are unbiased and consistent for both regression and dispersion parameters. In addition, the nature of the deviance surfaces suggests that these parameters are also orthogonal for most of the parameter space, which is advantageous for interpretation, inference, and computational efficiency. Study of the distribution's properties, through a consideration of dispersion, zero-inflation, and heavy tail indexes, together with the results of data analyses show the flexibility over standard approaches. The computational routines for fitting the original and reparameterized versions of the COM-Poisson regression model and data sets are available in appendix.

Keywords: COM-Poisson, Count data, Likelihood inference, Overdispersion, Underdispersion.

3.1 Introduction

The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution that includes the Poisson and geometric distributions as special cases, as well as the Bernoulli distribution as a limiting case. It can deal with both under- and overdispersed count data. Some recent applications of the COM-Poisson distribution include Lord, Geedipally, and Guikema (2010) for the analysis of traffic crash data, Sellers and Shmueli (2010) for the modelling of airfreight breakage and book purchases, Huang (2017) on the analysis of attendance data, takeover bids and cotton boll counts, and Chatla and Shmueli (2018) to model counts from bike sharing systems. Theoretical results and approximations derived for this distribution are discussed by Shmueli et al. (2005), Daly and Gaunt (2016), and Gaunt et al. (2017). The main disadvantage of the COM-Poisson regression model as presented in Sellers and Shmueli (2010) is that its location parameter does not correspond to the expectation of the distribution. This complicates the interpretation of regression models and means that they are not comparable with standard approaches, such as the Poisson and negative binomial regression models. In order to avoid this issue, Huang (2017) proposed a mean-parametrization of the COM-Poisson distribution. In his approach the mean parameter is obtained as the solution of a non-linear equation defined as an infinite sum, which is computationally demanding and liable to numerical problems.

The main goal of this chapter is to propose a novel COM-Poisson parametrization based on the mean approximation presented by Shmueli et al. (2005). In this parametrization, the probability mass function is written in terms of μ and ϕ, where μ is the approximate expectation and ϕ is a dispersion parameter. In contrast to the original parametrization, the proposed parametrization leads to regression coefficients directly associated (approximately) with the expectation of the response variable, as is usual in the context of generalized linear models. Consequently, the resulting regression coefficients are comparable to those obtained by standard approaches, such as the Poisson and negative binomial regression models. Furthermore, our novel COM-Poisson parametrization is computationally simpler than the strategy proposed by Huang (2017), since it does not require any numerical method for solving non-linear equations. Also, we show that attractive properties like the orthogonality (Ross 1970; Cox and Reid 1987) between dispersion and regression parameters and consistency and asymptotic normality of the maximum likelihood estimators are retained.

This chapter is organized as follows. In Section 3.2 we present the COM-Poisson distribution and the approach proposed by Huang (2017). The newly proposed reparametrization is considered in Section 3.3, along with assessment of the moment approximation, and a study of distribution's properties. In Section 3.4 we present estimation and inference for the novel COM-Poisson regression model in a likelihood framework. The properties of the maximum likelihood estimators and their approximate orthogonality are assessed in Section 3.5 through simulation studies. We illustrate the application of the new COM-Poisson regression model with the analysis of three data sets. We provide an `R` implementation for the COM-Poisson and reparameterized COM-Poisson regression models, together with the analyzed data sets, in the appendix and online material[1].

3.2 Background

The COM-Poisson distribution generalizes the Poisson distribution in terms of the ratio between the probabilities of two consecutive events by adding an extra dispersion parameter (Sellers and Shmueli 2010). Let Y be a COM-Poisson random variable, then

$$\frac{\Pr(Y = y - 1)}{\Pr(Y = y)} = \frac{y^\nu}{\lambda}$$

while for the Poisson distribution this ratio is y/λ corresponding to $\nu = 1$. This allows the COM-Poisson distribution to deal with non-equidispersed count data. The probability mass function of the COM-Poisson distribution is given by

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \qquad y = 0, 1, 2, \dots, \tag{3.1}$$

where $\lambda > 0$, $\nu \geq 0$ and $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j/(j!)^\nu$ is a normalizing constant that depends on both parameters.

The $Z(\lambda, \nu)$ series diverges theoretically only when $\nu = 0$ and $\lambda \geq 1$, but numerically, for small values of ν combined with large values of λ, the sum is so huge it results in overflow. Table 3.1 shows the values of the normalizing constant based on one thousand terms in the summation, that is $\sum_{j=0}^{1000} \lambda^j/(j!)^\nu$, for different values of λ and ϕ.

In the first line of Table 3.1, we have mathematically divergent series, because $\sum_{j=0}^{\infty} \lambda^j$ is divergent when $\lambda \geq 1$. In other cases the series diverges numerically, due to the computational

[1] Available on http://www.leg.ufpr.br/~eduardojr/papercompanions

Table 3.1. Values for $Z(\lambda, \nu)$ normalizing constant (computed numerically with 1000 terms in the summation) for values of λ (0.5 to 50) and ϕ (0 to 1)

$\boldsymbol{\nu}$	$\boldsymbol{\lambda}$ 0.5	1	5	10	30	50
0	2.00	divergent*	divergent*	divergent*	divergent*	divergent*
0.1	1.92	7.64	divergent**	divergent**	divergent**	divergent**
0.2	1.86	5.25	3.17e+273	divergent**	divergent**	divergent**
0.3	1.81	4.32	1.60e+29	2.54e+282	divergent**	divergent**
0.4	1.77	3.80	4.71e+10	1.33e+56	divergent**	divergent**
0.5	1.74	3.47	1.34e+06	3.67e+22	3.32e+196	divergent**
0.6	1.72	3.23	2.05e+04	4.99e+12	1.73e+76	4.63e+177
0.7	1.70	3.06	2.37e+03	3.69e+08	4.93e+39	6.93e+81
0.8	1.68	2.92	6.49e+02	2.70e+06	5.09e+24	3.43e+46
0.9	1.66	2.81	2.74e+02	1.47e+05	1.80e+17	2.19e+30
1	1.65	2.72	1.48e+02	2.20e+04	1.07e+13	5.18e+21

divergent* denotes a mathematically divergent series; and divergent** a numerically divergent series.

storage limitation. For both forms of divergence it is impossible to compute COM-Poisson probabilities, therefore, this places a restriction on the feasible parameter space.

An undesirable feature of the COM-Poisson distribution is that the moments cannot be obtained in closed form. Shmueli et al. (2005) and Sellers and Shmueli (2010) using an asymptotic approximation for $Z(\lambda, \nu)$, showed that the expectation and variance of the COM-Poisson distribution can be approximated by

$$\mathrm{E}(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \qquad \text{and} \qquad \mathrm{Var}(Y) \approx \frac{\lambda^{1/\nu}}{\nu}\,, \tag{3.2}$$

with greatest accuracy for $\nu \leq 1$ or $\lambda > 10^{\nu}$. The authors also argue that the mean-variance relationship can be approximated as $\mathrm{Var}(Y) \approx \mathrm{E}(Y)/\nu$. In Section 3.3, we assess the accuracy of these approximations.

The COM-Poisson regression model was proposed by Sellers and Shmueli (2010), using this original parametrization. Specifically, the COM-Poisson regression model is defined as $\log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ and the relationship between $\mathrm{E}(Y_i)$ and $\boldsymbol{x}_i$ is modelled indirectly. Huang (2017) shows how to model directly the expectation of the COM-Poisson distribution in a suitable reparametrization. From Equation (3.1), Huang notes that the parameter λ is given, as a function of μ and ν, by the solution to

$$\sum_{j=0}^{\infty} (j - \mu) \frac{\lambda^j}{(j!)^{\nu}} = 0\,.$$

This allows Huang to define a mean-parametrized COM-Poisson regression model with $\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ giving a direct relationship between μ and the covariates $\boldsymbol{x}$. In this article, we propose an alternative approximate mean-parametrization of the COM-Poisson distribution in order to avoid the limitations of the original parametrization and the numerical complexity of the Huang's approach.

3.3 Reparametrized COM-Poisson regression model

The proposed reparametrization of COM-Poisson models is based on the mean approximation (3.2). We introduce a new parameter μ, using this approximation,

$$\mu = h_\nu(\lambda) = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad \Rightarrow \quad \lambda = h_\nu^{-1}(\mu) = \left(\mu + \frac{(\nu-1)}{2\nu}\right)^\nu . \tag{3.3}$$

The dispersion parameter is taken on the log scale for computational convenience, thus $\phi = \log(\nu)$, $\phi \in \mathbb{R}$. The interpretation of ϕ is the same as the ν, but simply on another scale. For $\phi < 0$ and $\phi > 0$ we have the over- and underdispersion, respectively. When $\phi = 0$, we have the Poisson distribution as a special case.

In order to assess the accuracy of the moment approximations (Equation (3.2)), Figure 3.1 presents the errors for (a) expectation ($\mu - \mathrm{E}(Y)$) and (b) variance ($\mu \exp(-\phi) - \mathrm{Var}(Y)$). In both cases $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$ were computed numerically using 500 terms[2]. The dotted lines represent the border between the regions $\nu \leq 1$ and $\lambda > 10^\nu$, in the μ and ϕ scale.

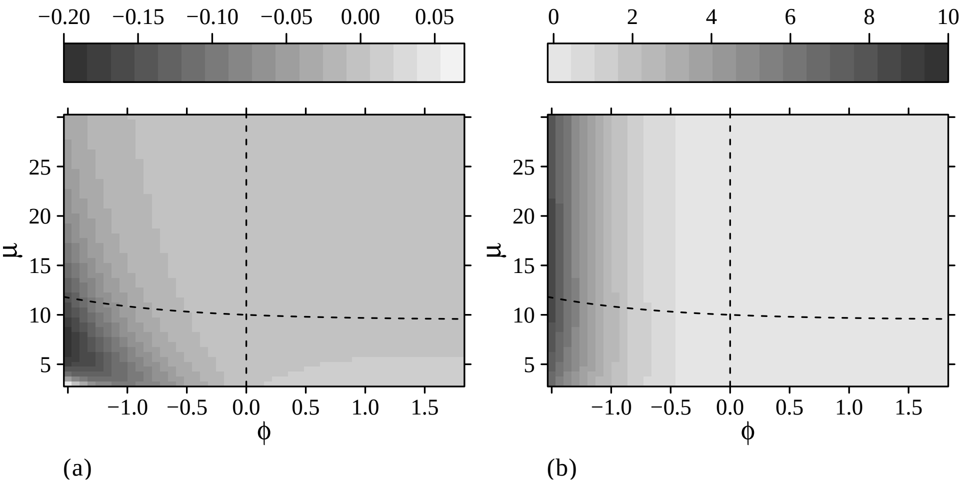

Figure 3.1. Squared errors for the approximation of the (a) expectation and (b) variance. Dotted lines represent the restriction for suitable approximations given by (Shmueli et al. 2005).

The results in Figure 3.1 show that the mean approximation is accurate, the largest (absolute) error, for the parameter values evaluated, is -0.197. For the variance approximation, the largest error was 8.4 and it occurs for negative values of ϕ. Interestingly, the errors for the variance are larger for negative values of ϕ and present no clear relation with μ, as opposed to the regions gives by Shmueli et al. (2005) ($\phi \leq 0$ and $\mu > 10 - (\exp(\phi) - 1)/(2\exp(\phi))$).

The results presented in Figure 3.1(a) support the proposed reparametrization. Replacing λ and ν as functions of μ and ϕ in Equation (3.1), the reparametrized distribution takes the form

$$\Pr(Y = y \mid \mu, \phi) = \left(\mu + \frac{e^\phi - 1}{2e^\phi}\right)^{ye^\phi} \frac{(y!)^{-e^\phi}}{Z(\mu, \phi)}, \qquad y = 0, 1, 2, \dots , \tag{3.4}$$

where $\mu > 0$. We denote this reparameterized distribution as COM-Poisson$_\mu$. In Figure 3.2, we show the shapes of the COM-Poisson$_\mu$ distribution.

[2] $\mathrm{E}(Y) \approx \sum_{y=0}^{500} y \Pr(Y = y)$ and $\mathrm{Var}(Y) \approx \sum_{y=0}^{500} [y - \mathrm{E}(Y)]^2 \Pr(Y = y)$.

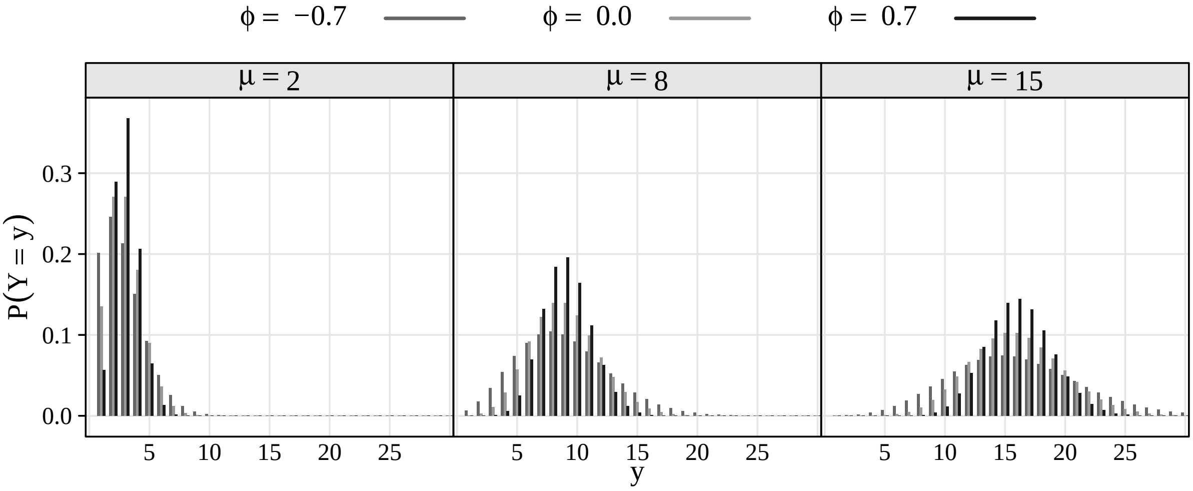

Figure 3.2. Shapes of the COM-Poisson distribution for different parameter values.

The constraint $\mu > 0$ imposes an undesirable restriction in the original parameter space, $\lambda^\nu > (\nu - 1)/(2\nu)$. However, this restricted region is related to small underdispersed counts (averages smaller than 0.1 and $\nu > 1$) a parameter region unlikely to be of interest in practice.

In order to explore the flexibility of the COM-Poisson model to deal with real count data, we compute indexes for dispersion (DI), zero-inflation (ZI) and heavy-tail (HT), which are respectively given by

$$\text{DI} = \frac{\text{Var}(Y)}{\text{E}(Y)}, \quad \text{ZI} = 1 + \frac{\log \Pr(Y = 0)}{\text{E}(Y)} \quad \text{and} \quad \text{HT} = \frac{\Pr(Y = y + 1)}{\Pr(Y = y)} \quad \text{for} \quad y \to \infty.$$

These indexes are defined in relation to the Poisson distribution. Thus, the dispersion index indicates overdispersion for $\text{DI} > 1$, underdispersion for $\text{DI} < 1$ and equidispersion for $\text{DI} = 1$. The zero-inflation index indicates zero-inflation for $\text{ZI} > 0$, zero-deflation for $\text{ZI} < 0$ and no excess of zeros for $\text{ZI} = 0$. Finally, the heavy-tail index indicates a heavy-tail distribution for $\text{HT} \to 1$ when $y \to \infty$. These indexes are discussed by Bonat et al. (2018) to study the flexibility of Poisson-Tweedie distribution, and Puig and Valero (2006) to describe count distributions in general.

Regarding the COM-Poisson$_\mu$ distribution, in Figure 3.3 we present the relationship between (a) mean and variance, (b–c) the dispersion and zero-inflation indexes for different values of μ and ϕ, and (d) the heavy-tail index for $\mu = 25$ and different values of y and ϕ. Figure 3.3 shows that the indexes are slightly dependent on the expected values and tend to stabilize for large values of μ. Consequently, the mean and variance relationship Figure 3.3(a) is proportional to the dispersion parameter ϕ. In terms of moments, this leads to a specification indistinguishable from the quasi-Poisson regression model. The dispersion indexes in Figure 3.3(b) show that the distribution is suitable to deal to dispersed counts, of course. For the parameter values evaluated the largest DI was 4.21 and smallest was 0.168. Figure 3.3(c) shows the COM-Poisson can handle a limited amount of zero-inflation, in cases of overdispersion ($\phi < 0$). On the other hand, for $\phi > 0$ (underdispersion) this distribution is suitable to deal with

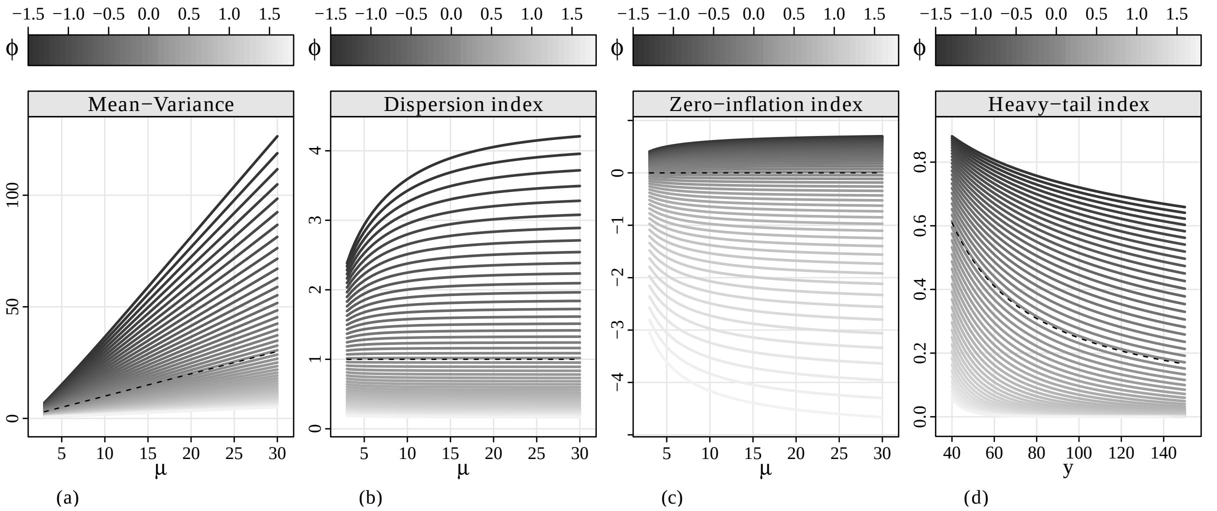

Figure 3.3. Indexes for the COM-Poisson distribution. (a) Mean and variance relationship, (b–d) dispersion, zero-inflation and heavy-tail indexes for different parameter values. Dotted lines represents the special case of the Poisson distribution.

zero-deflated counts. Heavy-tail indexes in Figure 3.3(d) indicate the distribution is in general a light-tailed distribution, i.e. HT → 0 for $y \to \infty$.

3.4 Estimation and inference

In this section we describe estimation and inference for the two forms of the COM-Poisson regression model based on the maximum likelihood method. Inference can be done using the standard machinery of likelihood inference, including likelihood ratio tests for model comparison and Wald-tests for testing individual (or groups of) parameters. The log-likelihood function for a set of independent observations y_i, $i = 1, 2, \dots, n$ from the COM-Poisson$_\mu$ distribution has the following form,

$$\ell = \ell(\boldsymbol{\beta}, \phi \mid \boldsymbol{y}) = e^{\phi} \left[\sum_{i=1}^{n} y_i \log \left(\mu_i + \frac{e^{\phi} - 1}{2e^{\phi}} \right) - \sum_{i=1}^{n} \log(y_i!) \right] - \sum_{i=1}^{n} \log[Z(\mu_i, \phi)], \tag{3.5}$$

where $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, with $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \dots, x_{ip})$ is a vector of known covariates for the ith observation, and $(\boldsymbol{\beta}, \phi) \in \mathbb{R}^{p+1}$. The normalizing constant $Z(\mu_i, \phi)$ is given by

$$Z(\mu_i, \phi) = \sum_{j=0}^{\infty} \left[\left(\mu_i + \frac{e^{\phi} - 1}{2e^{\phi}} \right)^{je^{\phi}} \frac{1}{(j!)^{e^{\phi}}} \right]. \tag{3.6}$$

The evaluation of the log-likelihood function for each observation involves the computation of the infinite series (3.6). Thus, the fitting procedure is computationally expensive for regions of the parameter space where the convergence of the infinite sum is slow.

Parameter estimation requires the numerical maximization of Equation (3.5). Since the derivatives of ℓ cannot be obtained in closed forms, we compute them numerically. We use the `BFGS` algorithm (Nocedal and Wright 2006) as implemented in the function `optim()` in `R` (R Core Team 2017). Standard errors for the regression coefficients are obtained based on the observed information matrix $\mathcal{I}(\boldsymbol{\theta})$, where $\mathcal{I}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta})$ (Hessian matrix) is computed numerically by central finite differences. Standard errors for $\hat{\eta}_i = \log(\hat{\mu}_i)$ and hence confidence intervals are obtained by using the delta method (Pawitan 2001, p.89).

In the original parametrization, parameter estimation is analogous to that presented for the COM-Poisson$_\mu$ distribution, however, it uses (3.5) re-expressed in terms of λ. Here, even for the standard COM-Poisson distribution, the dispersion parameter is taken on the log scale to avoid numerical issues.

For the applications we also fitted the quasi-Poisson model (Wedderburn 1974) as a baseline model. This approach is based only on a second-moment assumption and without specific underlying probablity model is less restrictive. In this model the variance of the response variable is specified by an additional dispersion parameter σ, with $\text{Var}(Y_i) = \sigma\mu_i$. These models are fitted in R using the function `glm(..., family = quasipoisson)`.

3.5 Simulation study

In this section, we report a simulation study to assess the properties of the maximum likelihood estimators, the approximate parameter orthogonality of the reparametrized model, as well as the flexibility of the COM-Poisson regression model to deal with non-equidispersed count data.

We considered average counts varying from 3 to 27 arising from a regression model with a continuous and a categorical covariate. The continuous covariate ($\boldsymbol{x}_1$) was generated as a linearly increasing sequence from 0 to 1 with length equal to the sample size. Similarly, the categorical covariate ($\boldsymbol{x}_2$) was generated as a sequence of three values each one repeated $n/3$ times (rounding up when required), where n denotes the sample size. The parameter μ of the reparametrized COM-Poisson random variable is given by $\boldsymbol{\mu} = \exp(\beta_0 + \beta_1\boldsymbol{x}_1 + \beta_{21}\boldsymbol{x}_{21} + \beta_{22}\boldsymbol{x}_{22})$, where $\boldsymbol{x}_{21}$ and $\boldsymbol{x}_{22}$ are dummy representing the levels of $\boldsymbol{x}_2$. The regression coefficients were fixed at the values, $\beta_0 = 2$, $\beta_1 = 0.5$, $\beta_{21} = 0.8$ and $\beta_{22} = -0.8$.

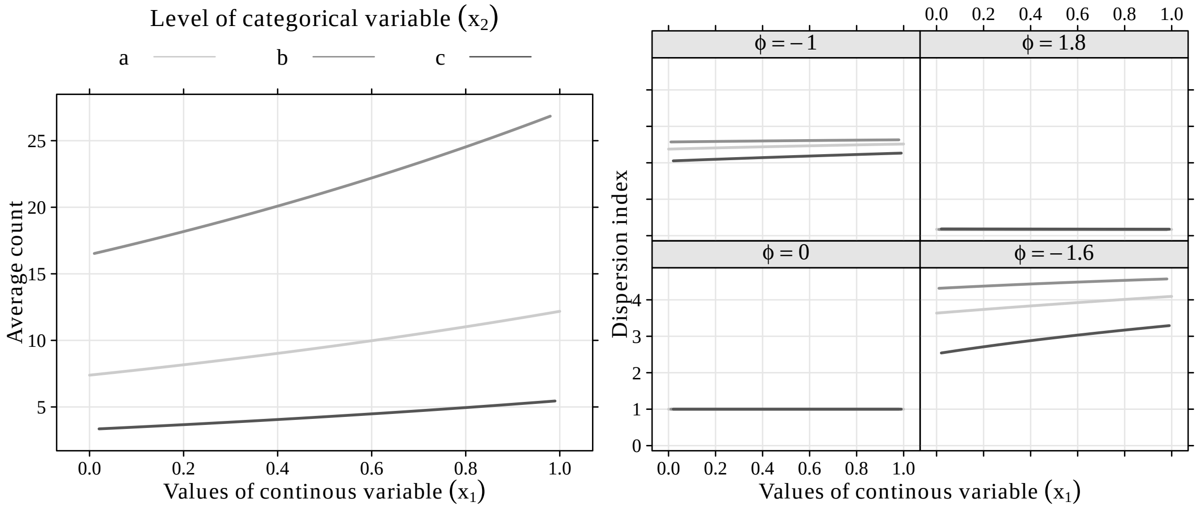

Figure 3.4. Average counts (left) and dispersion indexes (right) for each scenario considered in the simulation study.

We designed four simulation scenarios by considering different values of the dispersion parameter $\phi = -1.6, -1.0, 0.0$ and 1.8. Thus, we have strong and moderate overdispersion, equidispersion, and underdispersion, respectively. Figure 3.4 shows the variation of the average counts (left) and dispersion index (right) for each value of the dispersion parameter considered in the simulation study. These configurations allow us to assess the properties of the maximum likelihood estimators not only in extreme situations, such as high counts and low dispersion, and low counts and high dispersion, but also in the standard case of equidispersion.

In order to check the consistency of the estimators we considered four different sample sizes: 50, 100, 300 and 1000; generating 1000 data sets in each case. For sample sizes 50 and 100, we have 29 and 8 simulations of the 1000 where the fitting algorithm did not converge. These non-convergence situations occurred for $\phi = -1.6$. In Figure 3.5, we show the bias of the estimators for each simulation scenario (combination between values of the dispersion parameter and samples sizes) along with the confidence intervals calculated as average bias plus and minus 1.96 times the average standard error. The scales are standardized for each parameter by dividing the average bias by the average standard error obtained for the sample of size 50.

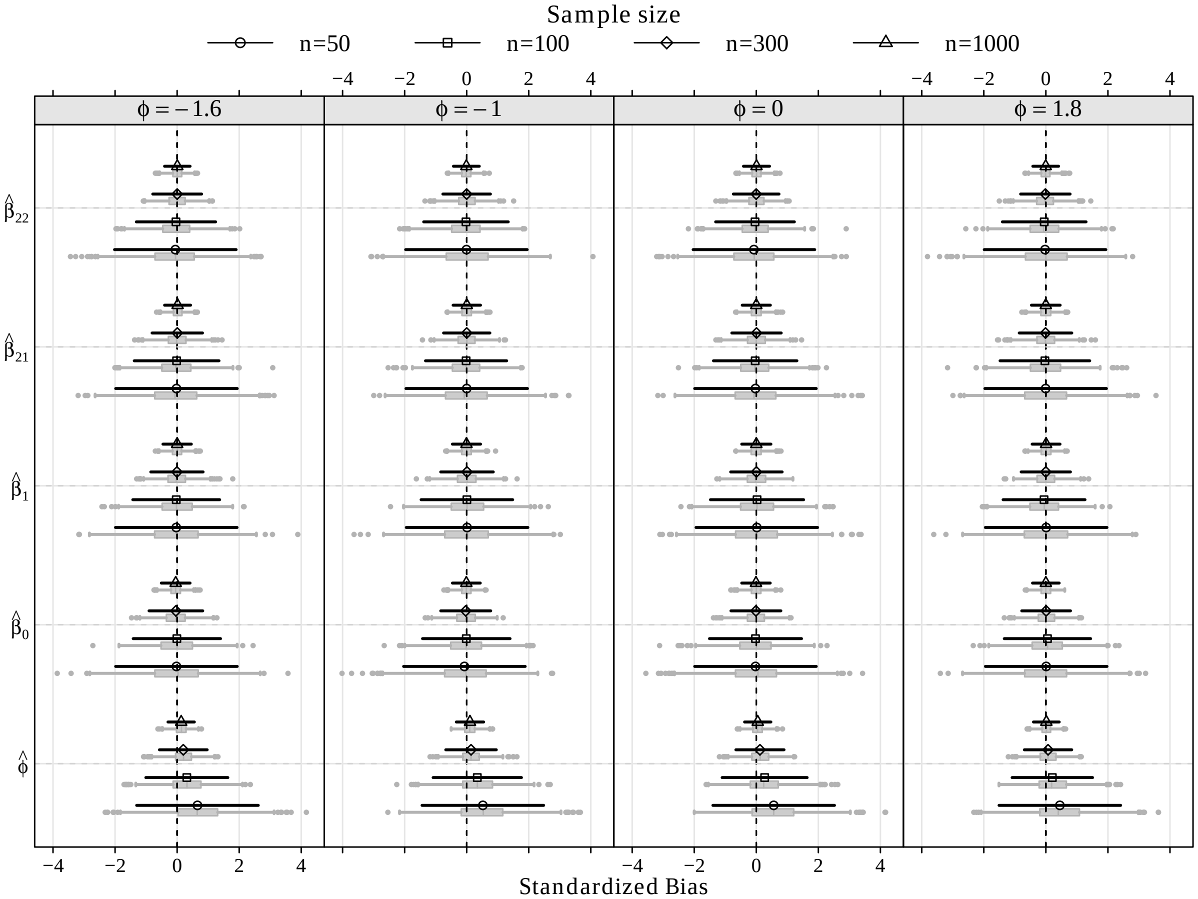

Figure 3.5. Distributions of standardized bias (gray box-plots) and average with confidence intervals (black segments) by different sample sizes and dispersion levels.

The results in Figure 3.5 show that for all dispersion levels, both the average bias and standard errors tend to 0 as the sample size increases. Thus the estimators for the regression parameters are unbiased, consistent and their empirical distributions are symmetric. For the dispersion parameter, the estimator is asymptotically unbiased; in small samples the parameter is overestimated and the empirical distribution is slightly right-skewed.

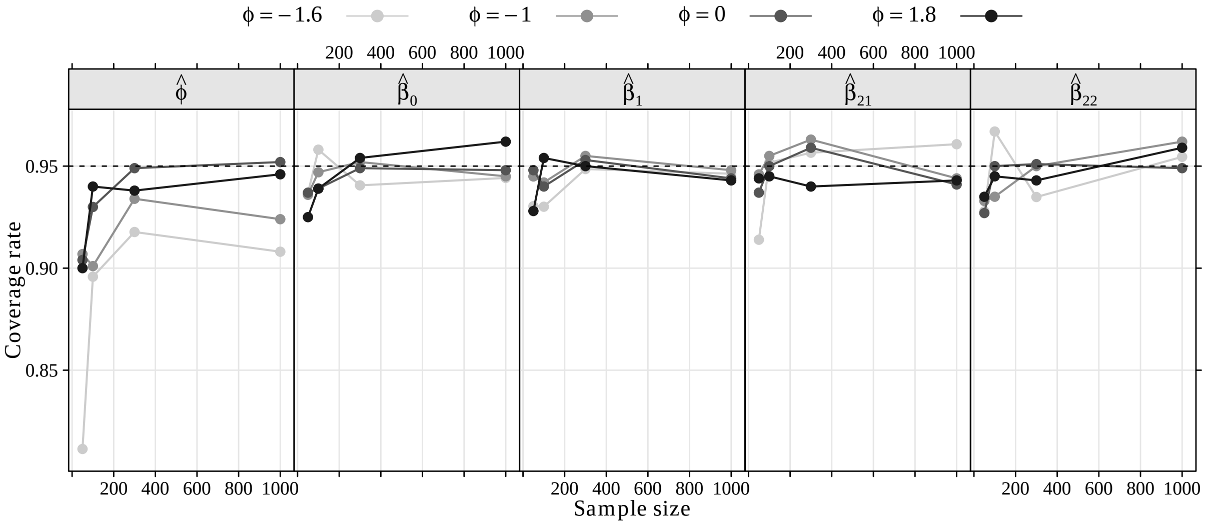

Figure 3.6. Coverage rate based on confidence intervals obtained by quadratic approximation for different sample sizes and dispersion levels.

Figure 3.6 presents the empirical coverage rate of the asymptotic confidence intervals. The results show that for the regression parameters the empirical coverage rates are close to the nominal level of 95% for sample sizes greater than 100 and all simulation scenarios. For the dispersion parameter the empirical coverage rates are slightly lower than the nominal level; however, they become closer to the nominal level for large samples. The worst scenario is when we have a small sample size and strong overdispersed counts.

To check the orthogonality property we compute the covariance matrix between maximum likelihood estimators $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\phi})$, obtained from the observed information matrix, $\text{Cov}(\hat{\boldsymbol{\theta}}) = \mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$. Figure 3.7 shows the correlation between the regression and dispersion parameter estimators for each simulation scenario, on the correlation scale. The correlations are generally close to zero in all cases suggesting the orthogonality property for the reparametrized model. Interestingly, results in the first panel show that $\text{Corr}(\hat{\beta}_{22}, \hat{\phi})$ is not very close to zero (values between -0.4 and 0.2) for strong overdispersion ($\phi = -1.6$).

To illustrate the orthogonality, Figure 3.8 displays contour plots of the deviance surfaces for four simulated data sets of size 1000 with $\mu = 5$ and different values of the dispersion parameters. The shapes of the deviance function show that the proposed parametrization is better for both computation and asymptotic (normal-based) inference. Furthermore, it is interesting to note that the deviance function shape under strong overdispersion ($\phi = -1.6$) is not as well behaved as the others; this is due to the slightly inaccuracy of the mean approximation (Figure 3.1(a)). Although the reparametrized model is always valid, μ and ϕ are orthogonal only when μ is the expectation of the distribution, in other words, when the approximation $\lambda - (\nu - 1)/(2\nu)$ is accurate. This also explains the results of $\text{Corr}(\hat{\beta}_{22}, \phi)$ in the first panel of Figure 3.7, since β_{22} is negative and associated with low counts.

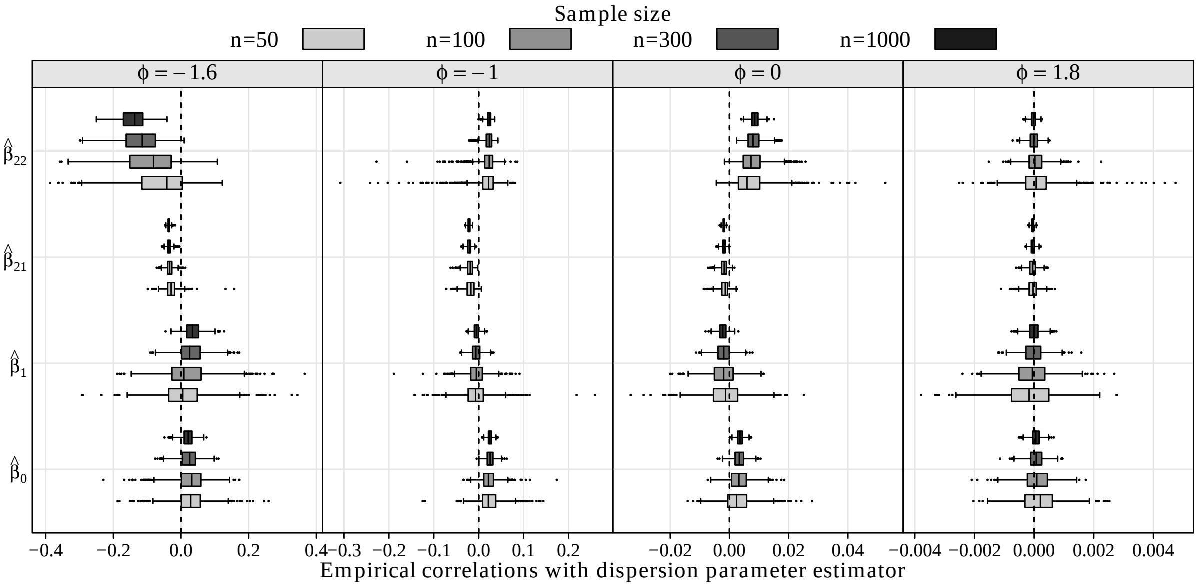

Figure 3.7. Empirical correlations between regression and dispersion parameters by different sample sizes and dispersion levels.

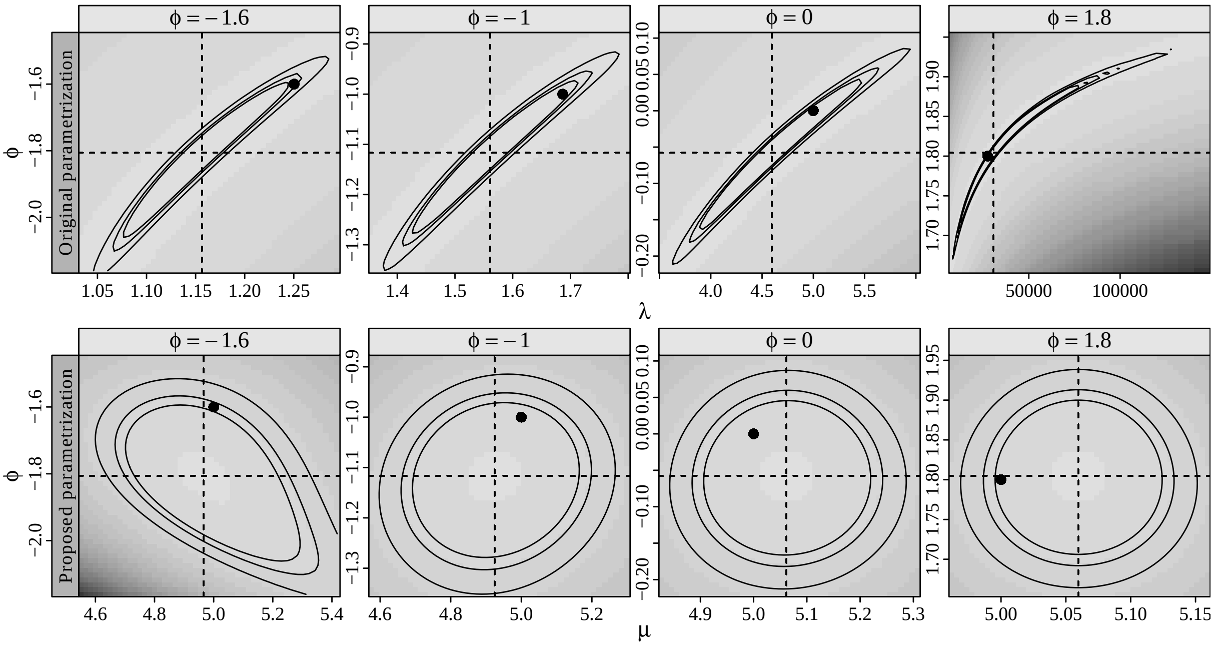

Figure 3.8. Deviance surfaces contour plots under original and proposed parametrization for four simulated data sets of size 1000 with different dispersion parameters. The ellipses are confidence regions (90, 95 and 99%), dotted lines are the maximum likelihood estimates, and points are the real parameters used in the simulation.

3.6 Case studies

In this section, we presented three illustrative examples of count data analysis. For the analyses as alternative models we considered the standard Poisson regression model, the COM-Poisson model in the two forms (original and new parametrization), and the quasi-Poisson regression model. The data sets and `R` code for their analysis are available in the appendix.

3.6.1 Artificial defoliation in cotton phenology

Following Zeviani et al. (2014), the linear predictor is given by $\eta_{ij} = \beta_0 + \beta_{1j}\texttt{def}_i + \beta_{2j}\texttt{def}_i^2$, where $\eta_{ij} = \log(\lambda_{ij})$ for the original COM-Poisson and $\eta_{ij} = \log(\mu_{ij})$ for COM-Poisson$_\mu$. The parameter μ_{ij} is the (approximated) expected number of cotton bolls for the ith defoliation level ($i =$ 1: 0%, 2: 25%, 3: 50%, 4: 75% e 5: 100%) and jth growth stage ($j =$ 1: vegetative, 2: flower bud, 3: blossom, 4: boll, 5: boll open), i.e. we have a different second order effect of defoliation at each growth stage. The parameter estimates and goodness-of-fit measures for the Poisson, COM-Poisson, COM-Poisson$_\mu$, and quasi-Poisson regression models are presented in Table 3.2.

Table 3.2. Parameter estimates (Est) and ratio between estimate and standard error (SE) for the four model strategies for the analysis of the cotton experiment.

	Poisson		COM-Poisson		COM-Poisson$_\mu$		Quasi-Poisson	
	Est	Est/SE	Est	Est/SE	Est	Est/SE	Est	Est/SE
σ	–	–	–	–	–	–	0.2410	–
ϕ	–	–	1.5847	12.4166	1.5817	12.3922	–	–
β_0	2.1896	34.5724	10.8967	7.7594	2.1905	74.6397	2.1896	70.4205
β_{11}	0.4369	0.8473	2.0187	1.7696	0.4350	1.8194	0.4369	1.7260
β_{12}	0.2897	0.5706	1.3431	1.2109	0.2876	1.2227	0.2897	1.1622
β_{13}	−1.2425	−2.0581	−5.7505	−3.8858	−1.2472	−4.4202	−1.2425	−4.1921
β_{14}	0.3649	0.6449	1.5950	1.2975	0.3500	1.3280	0.3649	1.3135
β_{15}	0.0089	0.0178	0.0377	0.0346	0.0076	0.0324	0.0089	0.0362
β_{21}	−0.8052	−1.3790	−3.7245	−2.7754	−0.8033	−2.9613	−0.8052	−2.8089
β_{22}	−0.4879	−0.8613	−2.2647	−1.8051	−0.4858	−1.8499	−0.4879	−1.7544
β_{23}	0.6728	0.9892	3.1347	2.0837	0.6788	2.1349	0.6728	2.0149
β_{24}	−1.3103	−1.9477	−5.8943	−3.6567	−1.2875	−4.0951	−1.3103	−3.9672
β_{25}	−0.0200	−0.0361	−0.0901	−0.0755	−0.0189	−0.0740	−0.0200	−0.0736
LogLik	−255.803		−208.250		−208.398		–	
AIC	533.606		440.500		440.795		–	
BIC	564.718		474.440		474.735		–	

The results presented in Table 3.2 show that the goodness-of-fit measures (log-likelihood, AIC and BIC) are quite similar for the COM-Poisson and COM-Poisson$_\mu$ models. This suggests that the reparametrization does not change the model fit, as is to be expected. The Poisson model is clearly unsuitable, being overly conservative, due to the overestimated standard errors. The difference in terms of minus twice log-likelihood value from the Poisson to the COM-Poisson$_\mu$ model is 94.811 for one additional parameter, which confirms the significantly improved fit of the COM-Poisson$_\mu$ model. Finally, the estimated value of the dispersion parameter $\hat{\phi} = 1.582$ suggests substantial underdispersion.

Furthermore, results in Table 3.2 also show the advantage of the COM-Poisson$_\mu$ model, since the regression parameter estimates are quite similar to those obtained for the Poisson model, whereas estimates from the original COM-Poisson model are on a different and not easily interpreted scale. The ratios between estimates and their respective standard errors for the two COM-Poisson models are very close to ratios from the quasi-Poisson model. However, it is important to note that COM-Poisson model is a full parametric approach, i.e. there is a probability distribution associated to the counts, while the quasi-Poisson model is based only on second-moment assumptions.

Figure 3.9 presents the observed and fitted values with 95% confidence intervals as a function of the defoliation level for each growth stage. The fitted values are the same for the Poisson and COM-Poisson$_\mu$ models, however, the confidence intervals are larger for the Poisson model because of the inappropriate equidispersion assumption. The results from the COM-Poisson$_\mu$ model are consistent with those from the Gamma-count model (Zeviani et al. 2014), Poisson-Tweedie (Bonat et al. 2018) and the alternative parametrization of the COM-Poisson distribution proposed by Huang (2017). All of these models indicated underdispersion and significant effects of defoliation at the vegetative, blossom and boll growth stages.

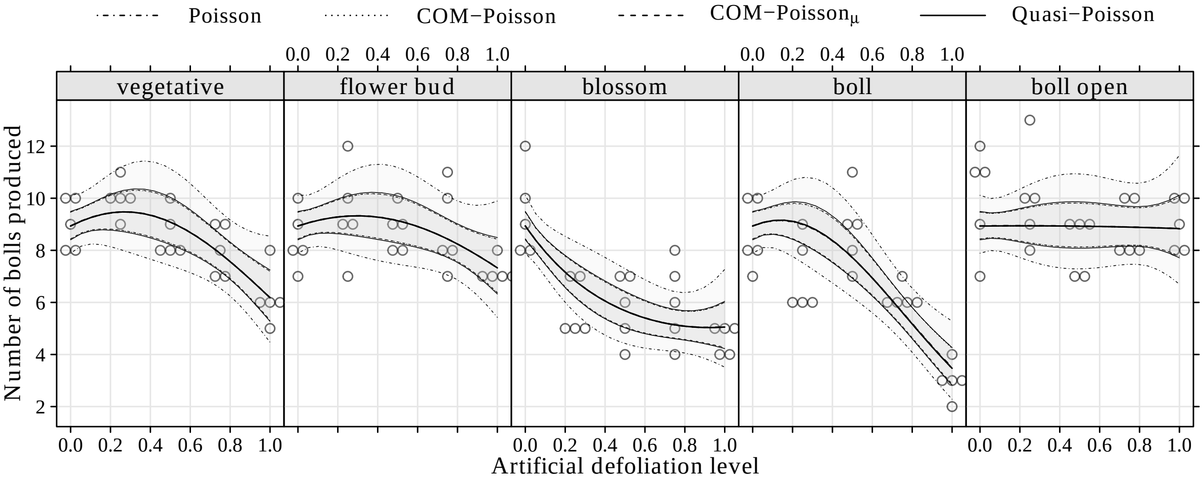

Figure 3.9. Scatterplots of the observed data and curves of fitted values with 95% confidence intervals as functions of the defoliation level for each growth stage.

In order to assess the relation between $\boldsymbol{\mu}$ and ϕ in the COM-Poisson$_\mu$ parametrization, Table 3.3 presents the empirical correlations between the regression and dispersion parameters, as computed by the asymptotic covariance matrix of the estimates, i.e. the inverse of the observed information. The correlations are practically zero for the COM-Poisson$_\mu$, however, with the original parametrization some correlations are quite large, in particular for the parameter β_0 (due to the effects parametrization in the linear predictor). This result explains the improved performance of the maximization algorithm in the new parametrization. It is also important to note that as initial regression parameter values for the `BFGS` algorithm are provided by the Poisson model, then in the COM-Poisson$_\mu$ model the initial values are very close to the maximum likelihood estimates and maximization effort is essentially only on the dispersion parameter ϕ. To compare the computational times for the two parametrizations we repeat the fitting 50 times.

In this example the COM-Poisson$_\mu$ fit was, on average, 38% faster than for the original version.

Table 3.3. Empirical correlations between $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ for the two parametrizations of COM-Poisson model fit to underdispersed data.

	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$
COM-Poisson	0.9952	0.2229	0.1526	−0.4895	0.1614	0.0043
COM-Poisson$_\mu$	0.0005	−0.0002	−0.0002	−0.0007	−0.0015	−0.0002
	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\beta}_{23}$	$\hat{\beta}_{24}$	$\hat{\beta}_{25}$	
COM-Poisson	−0.3496	−0.2276	0.2629	−0.4578	−0.0095	
COM-Poisson$_\mu$	0.0001	0.0002	0.0006	0.0018	0.0001	

3.6.2 Soil moisture and potassium doses on soybean culture

In this second example, based on the descriptive analysis (Section 2.2), we considered the following linear predictor

$$\eta_{ijk} = \beta_0 + \gamma_i + \tau_j + \beta_1 \texttt{K}_k + \beta_2 \texttt{K}_k^2 + \beta_{3j} \texttt{K}_k$$

where γ_i is the effect of ith block ($i = 1$: block II, 2: block III, 3: block IV and 4: block V), τ_j is the effect of jth moisture level ($j = 1$: 50% and 2: 62.5%), β_1 and β_2 give the baseline quadratic response over potassium levels `K` ($k = 1$: 0.0, 2: 0.3, 3: 0.6, 4: 1.2, 5: 1.8 100mg dm^{-3}) and β_{3j} is interaction of the first order potassium effect (`K`) for the jth moisture level (`umid`); The predictor η_{ijk} is $\log(\lambda_{ijk})$ for the original COM-Poisson and $\log(\mu_{ijk})$ for COM-Poisson$_\mu$. Table 3.4 presents the estimates, ratio between estimate and standard error, and goodness-of-fit measures for the alternative models.

Table 3.4. Parameter estimates (Est) and ratio between estimate and standard error (SE) for the four model strategies for the analysis of the soybean experiment.

	Poisson		COM-Poisson		COM-Poisson$_\mu$		Quasi-Poisson	
	Est	Est/SE	Est	Est/SE	Est	Est/SE	Est	Est/SE
σ	–	–	–	–	–	–	2.6151	–
ϕ	–	–	−0.7785	−4.7208	−0.7821	−4.7371	–	–
β_0	4.8666	144.2886	2.2320	6.0415	4.8666	97.7808	4.8666	89.2254
γ_1	−0.0194	−0.7302	−0.0089	−0.4939	−0.0194	−0.4950	−0.0194	−0.4516
γ_2	−0.0366	−1.3733	−0.0169	−0.9212	−0.0366	−0.9306	−0.0366	−0.8492
γ_3	−0.1056	−3.8890	−0.0486	−2.4223	−0.1056	−2.6338	−0.1056	−2.4049
γ_4	−0.0916	−3.2997	−0.0422	−2.1020	−0.0917	−2.2366	−0.0916	−2.0405
τ_1	0.1320	3.6471	0.0609	2.2949	0.1320	2.4715	0.1320	2.2553
τ_2	0.1243	3.4319	0.0573	2.1772	0.1243	2.3258	0.1243	2.1222
β_1	0.6160	11.0139	0.2839	4.7291	0.6161	7.4640	0.6160	6.8108
β_2	−0.2759	−10.2501	−0.1272	−4.5890	−0.2760	−6.9458	−0.2759	−6.3385
β_{31}	0.1456	4.2680	0.0670	2.6138	0.1456	2.8922	0.1456	2.6392
β_{32}	0.1648	4.8294	0.0759	2.8843	0.1648	3.2723	0.1648	2.9864
LogLik	−340.082		−325.241		−325.233		–	
AIC	702.164		674.482		674.467		–	
BIC	727.508		702.130		702.116		–	

The results in Table 3.4 show that the two parametrizations of the COM-Poisson model presented very similar goodness-of-fit measures and a better fit than the Poisson model. The minus twice difference between the log-likelihoods of the Poisson and COM-Poisson models was 29.697 on 1 degree of freedom, indicating that ϕ is significantly different from zero. The estimate of ϕ (-0.782) indicates overdispersion, corroborating the descriptive analysis. Concerning the regression parameters, the similarities between the models are analogous to those in the previous section. Both models show effects of block, potassium dose and moisture level, however the Poisson model indicates the effects with greater significance, because it does not take account of the extra variability and so underestimates standard errors.

The infinite sum $Z(\mu, \phi)$ in the cases of overdispersed count data requires a larger upper bound to reach convergence. Thus, in these cases the computation of the log-likelihood function is computationally expensive. For the data set considered here , the upper bound was fixed at 700. To reach convergence the `BFGS` algorithm evaluated the log-likelihood function 264 times when using the original parametrization of the COM-Poisson distribution and only 20 times with the new parametrization. In terms of computational time, for 50 repetitions, the proposed reparametrization was, on average, 110% faster to fit than the original one. This is probably due to the better behaviour of the log-likelihood function as well as more relevant initial values from the Poisson fit. In Table 3.5, we present the empirical correlations between the regression and dispersion parameter estimates. Again, the correlations are close to zero for the COM-Poisson$_\mu$ model, indicating the empirical orthogonality between μ and ϕ.

Table 3.5. Empirical correlations between $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ for the two parametrizations of COM-Poisson model fit to overdispersed data.

	$\hat{\beta}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\tau}_1$	$\hat{\tau}_2$
COM-Poisson	0.9952	0.2229	0.1526	−0.4895	0.1614	0.0043	−0.3496
COM-Poisson$_\mu$	0.0005	−0.0002	−0.0002	−0.0007	−0.0015	−0.0002	0.0001
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{31}$	$\hat{\beta}_{32}$			
COM-Poisson	−0.2276	0.2629	−0.4578	−0.0095			
COM-Poisson$_\mu$	0.0002	0.0006	0.0018	0.0001			

The observed and fitted counts for each moisture level with confidence intervals are shown in Figure 3.10. The fitted values are identical for the Poisson and COM-Poisson$_\mu$ models, leading to the same conclusions. On the other hand, confidence intervals for the Poisson model are narrower than those from the COM-Poisson$_\mu$, due to the inappropriate equidispersion assumption of the Poisson model. The confidence intervals from the COM-Poisson$_\mu$ and quasi-Poisson models are also very similar, which again shows the already highlighted similarity between these approaches, however only the COM-Poisson model$_\mu$ corresponds to a fully specified probability model.

3.6.3 Assessing toxicity of nitrofen in aquatic systems

This data set comes from an completely randomized experiment to measure the reproductive toxicity of the herbicide nitrofen on a species of zooplankton (Section 2.3). Here, we consider three models with linear predictors,

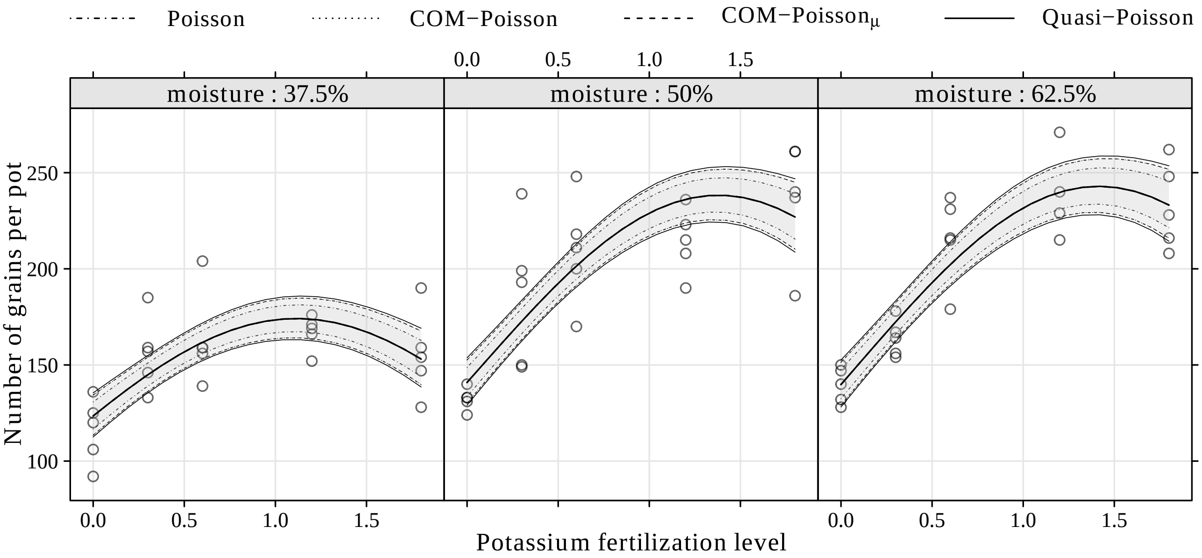

Figure 3.10. Observed bean seeds counts as functions of potassium doses and soil moisture levels with fitted curves and confidence intervals (95%).

$$\begin{aligned}
&\text{Linear:} && \eta_i = \beta_0 + \beta_1 \texttt{dose}_i,\\
&\text{Quadratic:} && \eta_i = \beta_0 + \beta_1 \texttt{dose}_i + \beta_2 \texttt{dose}_i^2 \text{ e}\\
&\text{Cubic:} && \eta_i = \beta_0 + \beta_1 \texttt{dose}_i + \beta_2 \texttt{dose}_i^2 + \beta_3 \texttt{dose}_i^3,
\end{aligned}$$

where $\eta_i = \log(\lambda_i)$ for the original COM-Poisson and $\eta_i = \log(\mu_i)$ for COM-Poisson$_\mu$.

Table 3.6. Model fit measures and comparisons between predictors and models fitted to the nitrofen data.

Poisson	$\#p$	ℓ	AIC	2(diff ℓ)	diff $\#p$	P($> \chi^2$)	
Preditor 1	2	−180.667	365.335	–	–	–	–
Preditor 2	3	−147.008	300.016	67.319	1	2.31E−16	–
Preditor 3	4	−144.090	296.180	5.835	1	1.57E−02	–
COM-Poisson	$\#p$	ℓ	AIC	2(diff ℓ)	diff $\#p$	P($> \chi^2$)	$\hat{\phi}$
Preditor 1	3	−167.954	341.908	–	–	–	−0.893
Preditor 2	4	−146.964	301.929	41.980	1	9.22E−11	−0.059
Preditor 3	5	−144.064	298.129	5.800	1	1.60E−02	0.048
COM-Poisson$_\mu$	$\#p$	ℓ	AIC	2(diff ℓ)	diff $\#p$	P($> \chi^2$)	$\hat{\phi}$
Preditor 1	3	−167.652	341.305	–	–	–	−0.905
Preditor 2	4	−146.950	301.900	41.405	1	1.24E−10	−0.069
Preditor 3	5	−144.064	298.127	5.773	1	1.63E−02	0.047
Quasi-Poisson	$\#p$	QDev	AIC	F	diff $\#p$	P($> F$)	$\hat{\sigma}$
Preditor 1	3	123.929	–	–	–	–	2.262
Preditor 2	4	56.610	–	60.840	1	5.07E−10	1.106
Preditor 3	5	50.774	–	5.659	1	2.16E−02	1.031

$\#p$, number of parameters; diff ℓ, difference in log-likelihoods; QDev, quasi-deviance, F, F statistics based on quasi-deviances; diff $\#p$, difference in number of parameters.

Table 3.6 summarizes the results of the fitted models and likelihood ratio tests comparing the sequence of predictors. All models indicate the significance of the cubic effect of the nitrofen concentration. Considering this predictor, there is evidence of equidispersed counts, the ϕ estimate of the COM-Poisson is close to zero and the σ estimate of the quasi-Poisson is close to one. It is interesting to note that if we omit the higher order effects the models show evidence of overdispersion — this exemplifies the discussion on the possible causes of overdispersion in Section 3.1. We also note that the quasi-Poisson approach, although robust to the equidispersion assumption, shows higher descriptive levels (p-values) than parametric models, that is, the tests under parametric models are more powerful than the ones under the quasi-Poisson model in the equidispersed case.

In Table 3.7, we present the estimates of the regression parameters for the cubic dose models. The interpretations are similar to the other case studies, however, here the Poisson model is also suitable for indicating the significance of the covariate effects. In addition, note that the parameter estimates of the original COM-Poisson model are comparable to the others models. This occurs because we are in the particular case, where $\phi \approx 0$, which implies $\lambda \approx \mu$.

Table 3.7. Parameter estimates (Est) and ratio between estimate and standard error (SE) for the four cubic dose models for the analysis of the nitrofen experiment.

	Poisson		COM-Poisson		COM-Poisson$_\mu$		Quasi-Poisson	
	Est	Est/SE	Est	Est/SE	Est	Est/SE	Est	Est/SE
β_0	3.4767	62.8167	3.6494	4.8499	3.4769	64.3083	3.4767	61.8599
β_1	−0.0860	−0.4328	−0.0914	−0.4475	−0.0879	−0.4523	−0.0860	−0.4262
β_2	0.1529	0.8634	0.1612	0.8783	0.1547	0.8938	0.1529	0.8502
β_3	−0.0972	−2.3978	−0.1021	−2.2294	−0.0976	−2.4635	−0.0972	−2.3612

Figure 3.11 shows the number of live off-spring observed and fitted curves along with confidence intervals for the different cubic dose models. The fitted values and confidence intervals are identical and have a complete overlap. This shows that the estimation of the extra dispersion parameter does not affect the estimation of the regression coefficients in the case of equidispersed counts.

Finally, in Table 3.8 we present the empirical correlations between the regression and dispersion parameter estimates. The results show that even in the special case ($\phi = 0$), the empirical correlations for the original COM-Poisson model are not zero. For the reparametrized model, as discussed in the previous sections, the correlations are practically null. The computational times for fifty repetitions of fit are similar; the average time to fit the COM-Poisson$_\mu$ and COM-Poisson models is 1.19 and 1.09 seconds, respectively.

Table 3.8. Empirical correlations between $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ for the two parametrizations of COM-Poisson model fit to equidispersed data.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
COM-Poisson	0.9972	−0.0771	0.1562	−0.4223
COM-Poisson$_\mu$	−0.0003	0.0023	−0.0029	0.0033

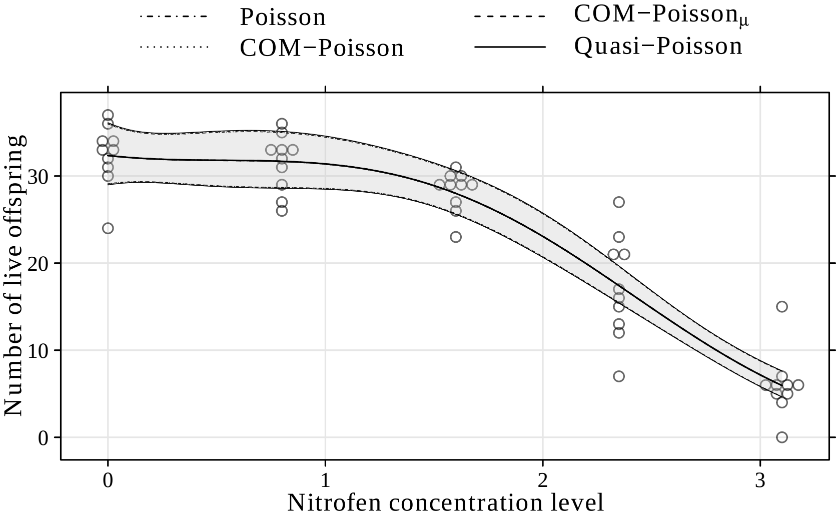

Figure 3.11. Number of live offspring observed for each nitrofen concentration level with fitted curves and 95% confidence intervals.

3.7 Concluding remarks

In this chapter, we presented and characterized a novel reparametrization of the COM-Poisson distribution and associated regression model. The reparametrization was based on a simple asymptotic approximation for the expectation of the COM-Poisson distribution. The main advantage of the proposed reparametrization is the simple interpretation of the regression coefficients in terms of the (approximate) expectation of the response variable as usual in the generalized linear models context. Thus, it is possible to compare directly the results from the COM-Poisson model with those from standard approaches, such as the Poisson and quasi-Poisson regression models. Furthermore, in the new parametrization the COM-Poisson distribution is indexed by μ and an extra dispersion parameter ϕ which our studies suggest are approximately orthogonal. Overall the approach is similar to Huang's (2017) parametrization but ours is simpler, because the μ used here is obtained from simple algebra.

The proposed parametrization is always valid, only the interpretation of the μ parameter, as the expectation of the distribution, depends on the accuracy of the approximation. We evaluated the accuracy of the approximations for the expectation and variance of the COM-Poisson distribution by considering quadratic approximation errors. We found that the mean approximation is slightly inaccurate only for strongly overdispersed counts with small averages. Therefore, otherwise, the mean approximation is reasonable and the regression coefficients can be interpreted in terms of expectation. In this situation the parameter estimates from the COM-Poisson model will be very close to those from the Poisson (and quasi-Poisson) fit, as was seen in the case studies, and this agreement can be considered as an indication of the reasonableness of the approximation.

We discuss the properties and flexibility of the distribution to deal with count data through considerations of dispersion, zero-inflation and heavy-tail indexes. A simulation study was used to assess the properties of the reparametrized COM-Poisson model and its ability to deal with different levels of dispersion, as well as the properties of the maximum likelihood estimators. The results of our simulation study suggested that the maximum likelihood estimators of the regression and dispersion parameters are unbiased and consistent. The empirical coverage rates of the confidence intervals computed based on the asymptotic distribution of the maximum likelihood estimators are close to the nominal level for sample sizes greater than 100. The worst case scenario is when we have a small sample size and strongly overdispersed counts. In general, we recommend the use of the asymptotic confidence intervals for computational simplicity, although of course bootstrap intervals could be obtained, but would involve extensive computation.

The three examples and data analyses have shown that the COM-Poisson regression model is a suitable choice to deal with generally dispersed count data, equi-, under- and overdispersed. The observed empirical correlation between the regression and dispersion parameter estimators and deviance surfaces suggest approximate orthogonality between μ and ϕ in the COM-Poisson$_\mu$ distribution. Thus, the computational procedure based on the proposed reparametrization is faster than that for the original parametrization.

In general, the results presented by the reparametrized COM-Poisson models were satisfactory and comparable to the conventional approaches. Therefore, its use in the analysis of count data is encouraged. The computational routines for fitting the original and reparametrized COM-Poisson regression models are available in the appendix.

References

Bonat, W. H., B. Jørgensen, C. C. Kokonendji, and J. Hinde (2018). “Extended Poisson-Tweedie: properties and regression model for count data”. In: *Statistical Modelling* 18.1, pp. 24–49.

Chatla, S. B. and G. Shmueli (2018). “Efficient estimation of COM-Poisson regression and a generalized additive model”. In: *Computational Statistics & Data Analysis* 121, pp. 71–89.

Cox, D. R. and N. Reid (1987). “Orthogonality and Approximate Conditional Inference (with discussion)”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 49.1, pp. 1–39.

Daly, F. and R. Gaunt (2016). “The Conway-Maxwell-Poisson distribution: Distributional theory and approximation”. In: *ALEA, Latin American Journal of Probability and Mathematical Statistics* 13, pp. 635–658.

Gaunt, R., S. Iyengar, A. Olde Daalhuis, and B. Simsek (2017). “An asymptotic expansion for the normalizing constant of the Conway-Maxwell-Poisson distribution”. In: *Annals of the Institute of Statistical Mathematics* to appear.

Huang, A. (2017). “Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts”. In: *Statistical Modelling* 17.6, pp. 1–22.

Lord, D., S. R. Geedipally, and S. D. Guikema (2010). “Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion”. In: *Risk Analysis* 30.8, pp. 1268–1276.

Nocedal, J. and S. J. Wright (2006). *Numerical optimization.* 2nd edition. Series in Operations Research. New York: Springer, p. 636. ISBN: 0387987932.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood.* New York: Oxford University Press.

Puig, P. and J. Valero (2006). “Count data distributions: some characterizations with applications”. In: *Journal of the American Statistical Association* 101.473, pp. 332–340.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

Ross, G. J. S. (1970). “The Efficient Use of Function Minimization in Non-Linear Maximum-Likelihood Estimation”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 19.3, pp. 205–221.

Sellers, K. F. and G. Shmueli (2010). “A flexible regression model for count data”. In: *Annals of Applied Statistics* 4.2, pp. 943–961.

Shmueli, G., T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright (2005). “A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.1, pp. 127–142.

Wedderburn, R. W. M. (1974). “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method”. In: *Biometrika* 61.3, p. 439.

Zeviani, W. M., P. J. Ribeiro Jr, W. H. Bonat, S. E. Shimakura, and J. A. Muniz (2014). “The Gamma-count distribution in the analysis of experimental underdispersed data”. In: *Journal of Applied Statistics* 41.12, pp. 2616–2626.

4 A REVIEW OF FLEXIBLE MODELS FOR DISPERSED COUNT DATA

ABSTRACT

In the analysis of count data often the equidispersion assumption is not suitable, hence the Poisson regression model is inappropriate. Several count distributions that can handle both under- and overdispersion have been proposed in the literature. However, a comparison of these distributions in the context of regression models has not yet been performed. In this chapter, we reviewed and compared the use of COM-Poisson, Gamma-count, discrete Weibull, generalized Poisson, double Poisson and Poisson-Tweedie distributions for practical data analysis. The main properties of the distributions are highlighted and compared considering the dispersion, zero-inflation, and heavy-tail indexes. The application of these models is illustrated with the analysis of two experimental data sets. The computational routines for fitting the models and the data sets are available in the appendix.

Keywords: COM-Poisson, Gamma-count, Generalized Poisson, Overdispersion, Poisson-Tweedie family, Underdispersion.

4.1 Introduction

Important advances for count data analysis have been reported in the literature. Among them, we mention, mainly, methods to model different levels of dispersion, since the standard Poisson model forces the equality between mean and variance, referred to as equidispersion, which rarely occurs in real data (Molenberghs, Verbeke, and Demétrio 2007). A comprehensive review of new methods for modeling count data can be found in the recent monographs by Winkelmann (2008), Cameron and Trivedi (2013), and Hilbe (2014).

The most common failure of the equidispersion assumption is the overdispersion (or extra-Poisson variability) when the variance exceeds the mean. Overdispersion count data has been well studied by the statistical community and there is a wide range of models to deal with. Hinde and Demétrio (1998), for example, discussed models and estimation algorithms for overdispersed discrete data, as the common choices negative binomial and quasi-Poisson. On the other hand, underdispersion, when the variance is smaller than the mean, is less often reported in the literature. However, in the last decade, it has been of increasing interest.

Shmueli et al. (2005) brought back the COM-Poisson model and supported several further research like as Lord, Guikema, and Geedipally (2008), Sellers and Shmueli (2010), Huang (2017), and Ribeiro Jr et al. (2018). Zeviani et al. (2014) discussed the analysis of experimental data based on the Gamma-count distribution. Klakattawi, Vinciotti, and Yu (2018) discussed the properties and application of the discrete Weibull distribution and Luyts et al. (2018) extended it to model longitudinal/clustered data. Zamani and Ismail (2012) presented an extension for the generalized Poisson regression model. Double Poisson distribution proposed by Efron (1986), can be used as the distribution for the response variable in general purposed class of models GAMLSS (generalized additive models for location, scale, and shape) (Rigby and Stasinopoulos 2005). Bonat et al. (2018) proposed the extended Poisson-Tweedie for handling under- and overdispersion count data based on moments of the Poisson-Tweedie family.

The main goal of this chapter is to compare several model strategies for analysis of dispersed count data, namely COM-Poisson, Gamma-count, discrete Weibull, generalized Poisson, double Poisson, and Poisson-Tweedie, in terms of characteristic indexes (dispersion, zero-inflation, and heavy-tail) and real applications. Related work can be found in Kokonendji (2014) who did a well-documented discussion of count statistical models, but without applications or model comparisons, and Sellers and Morris (2017) who discussed the causes of underdispersion and presented some count distributions with two applications.

This chapter is organized in six sections. The considered model strategies and their main properties are presented in Section 4.2. In Section 4.3, we present the characteristic indexes and use them to compare the models. The regression models and estimation are presented in Section 4.4. Section 4.5 is devoted to illustrating the application of the models for analysis of two datasets results from experimental studies and compare the results. Finally, discussion and direction for future work are given in Section 4.6. The computational routines used in this chapter are available in the appendix.

4.2 Background

In this section, we shall introduce some count distributions taking into account dispersed count data. Namely, we consider the COM-Poisson, Gamma-count, discrete Weibull, generalized Poisson, double Poisson, and Poisson-Tweedie distributions. We focus on the genesis of generating probability distributions and their main properties.

4.2.1 COM-Poisson distribution

The COM-Poisson distribution is an important member of the family of weighted Poisson distributions (WPD) (Del Castillo and Pérez-Casany 1998). The WPD family weights the Poisson probability function by a suitable function, allowing for a nonlinear decrease in ratios of successive probabilities. A random variable Y is a weighted Poisson distribution if its probability mass function can be written in the form

$$\Pr(Y = y) = \frac{\exp(-\lambda)\lambda^y}{y!}\frac{w(y)}{\mathrm{E}_\lambda[w(Y)]}, \quad y \in \mathbb{N},$$

where $\mathrm{E}_\lambda(\cdot)$ denotes the mean value with respect to the Poisson random variable with parameter $\lambda \geq 0$ and $w(\cdot)$ is a weight function. The weight function may depend on an extra parameter to ensure more flexibility.

The COM-Poisson distribution arises when $w(y) \equiv w(y,\nu) = (y!)^{1-\nu}$ for $\nu \geq 0$ (Kokonendji 2014). Its probability mass function takes the form

$$\Pr(Y = y) = \frac{\lambda^y \exp(-\lambda)}{(y!)^\nu \mathrm{E}_\lambda[(Y!)^{1-\nu}]} = \frac{\lambda^y}{(y!)^\nu \mathrm{Z}(\lambda,\nu)}, \quad \text{where} \quad \mathrm{Z}(\lambda,\nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}. \tag{4.1}$$

The series $\mathrm{Z}(\lambda,\nu)$ is a normalizing constant that cannot be expressed in closed form unless for special cases. The parameters λ and ν have no direct interpretation but can be seen in terms of the ratios of successive probabilities which is given, for the WPD family and particularly for the COM-Poisson distribution, by

$$\frac{\Pr(Y = y-1)}{\Pr(Y = y)} = \frac{y}{\lambda}\frac{w(y-1)}{w(y)} \overset{\text{CMP}}{\equiv} \frac{y^\nu}{\lambda},$$

a nonlinear function of y. As the dispersion parameter, for $0 < \nu < 1$ and $\nu > 1$ we have the over- and underdispersion, respectively. When $\nu = 1$ the Poisson distribution results as a special case. Another special case is the geometric distribution, when $\nu = 0$ and $\lambda < 1$ with $\Pr(Y = y) = \lambda^y(1 - \lambda)$. As a limiting case, the Bernoulli distribution arises when $\nu \to \infty$, with probability of success equal to $\lambda/(\lambda + 1)$.

The COM-Poisson also belongs to the family of two-parameter power series distributions (Johnson, Kemp, and Kotz 2005) and using the properties for this family, the mean and variance of COM-Poisson distributions are given by

$$\mathrm{E}(Y) = \lambda \frac{d}{d\lambda}[\log Z(\lambda, \nu)] \quad \text{and} \quad \mathrm{Var}(Y) = \lambda^2 \frac{d^2}{d\lambda^2}[\log Z(\lambda, \nu)] + \mathrm{E}(Y),$$

respectively, but cannot be solved in closed form. Therefore, despite the nice properties of COM-Poisson distribution, its major limitation is that the location parameter λ does not represent the expectation of the distribution and has no direct interpretation. Under this motivation, Huang (2017) and Ribeiro Jr et al. (2018) proposed new parametrizations of COM-Poisson distribution related to the expectation. The distribution is now indexed by the parameter μ, where μ is obtained by the solution for $\sum_{j=0}^{\infty}(j - \mu)\lambda^j/(j!)^\nu = 0$ in the Huang's proposal and $\mu = \lambda^{1/\nu} - (\nu - 1)/2\nu$ in the Ribeiro Jr's proposal. In this chapter, we use the Ribeiro Jr's parametrization for simplicity.

4.2.2 Gamma-count distribution

Another straightforward flexible distribution arises from the relationship between Poisson and exponential distribution. Following Winkelmann (1995), let $\tau_k > 0$, $k \in \mathbb{N}^*$, denote the waiting times between the $(k-1)$th and the kth event and ϑ_n, denote the arrival time of the nth event, so $\vartheta_n = \sum_{k=1}^{n} \tau_k$. Finally, denote Y_T the number of events within a $(0, T)$ interval. Following the definitions, we have

$$\begin{aligned} Y_T < y &\iff \vartheta_y \geq T \\ \Pr(Y_T < y) &= \Pr(\vartheta_y \geq T) = 1 - \mathrm{F}_{\vartheta_y}(T), \\ \Pr(Y_T = y) &= \Pr(Y_T < y) - \Pr(Y_T < y + 1) \\ \Pr(Y_T = y) &= \mathrm{F}_{\vartheta_y}(T) - \mathrm{F}_{\vartheta_{y+1}}(T), \end{aligned} \tag{4.2}$$

where $\mathrm{F}_{\vartheta_n}(T)$ is the cumulative density function of ϑ_n and T is the interval of the counting (*offset*). This process is called by renewal process (Cox 1962; Winkelmann 2008, p.54).

From the renewal process (4.2), Poisson distribution is obtained by assuming τ_k exponentially distributed. Consequently, ϑ_n drawn from an Erlang distribution (a special case of gamma distribution) and $\Pr(Y_T = y)$ has closed form

$$\Pr(Y_T = y) = \mathrm{F}_{\vartheta_y}(T) - \mathrm{F}_{\vartheta y+1}(T) = \sum_{j=0}^{y-1} \frac{\exp(-\lambda T)(\lambda T)^j}{j!} - \sum_{j=0}^{y} \frac{\exp(-\lambda T)(\lambda T)^j}{j!} = \frac{\exp(-\lambda T)(\lambda T)^j}{y!}.$$

Winkelmann (1995) proposed to use a more general distribution for the waiting times τ_k. The Gamma-count distribution assumes that τ_k is independently gamma distributed, i.e

$\tau_k \overset{iid}{\sim}$ Gamma(α, β). Due to the reproductive property of gamma random variables, $\vartheta_n \sim$ Gamma$(y\alpha, \beta)$. Thereby, the Gamma-count probability mass function takes the form

$$\Pr(Y_T = y) = \int_0^T \frac{\beta^{y\alpha} t^{y\alpha-1}}{\Gamma(y\alpha)\exp(\beta t)} dt - \int_0^T \frac{\beta^{(y+1)\alpha} t^{(y+1)\alpha-1}}{\Gamma[(y+1)\alpha]\exp(\beta t)} dt, \tag{4.3}$$

a difference between two gamma cumulative density functions, $G(y\alpha, \beta) - G((y+1)\alpha, \beta)$, where $G(a, b)$ is the cumulative function $\mathrm{F}_{\vartheta_y}(T)$ for the gamma variable with parameters a and b.

As in the COM-Poisson distribution, the moments cannot be obtained in closed form. However, Winkelmann (1995) using the results from Cox (1962), showed for increasing T, i.e. high counts, it holds that

$$Y_T \overset{asy}{\sim} \mathcal{N}\left(\frac{\beta T}{\alpha}, \frac{\beta T}{\alpha^2}\right),$$

thus the limiting variance-mean ratio equals a constant $1/\alpha$. Consequently, the Gamma-count distribution displays overdispersion for $0 < \alpha < 1$ and underdispersion for $\alpha > 1$.

In order to have a mean-type parameterization, we rewrite the Equation (4.3) in terms of $\kappa = \beta/\alpha$, thus the probability mass function is given by

$$\Pr(Y = y) = G(y\alpha, \kappa\alpha) - G((y+1)\alpha, \kappa\beta). \tag{4.4}$$

In this parametrization, the parameter κ is the expected value for the waiting times τ_k and not for the counts, unless $\alpha = 1$. However, for large T, the asymptotic distribution of Y holds and the expected value of Y is equal to κT.

4.2.3 Discrete Weibull distribution

The discrete Weibull distribution comes from a discretization process of the continuous Weibull distribution (Nakagawa and Osaki 1975). Let T be a continuous Weibull distributed random variable with parameters $\lambda > 0$ and $\rho > 0$. The density function and cumulative density function are, respectively, given by

$$f_T(t) = \lambda\rho t^{\rho-1}\exp(-\lambda t^\rho) \quad \text{and} \quad F_T(t) = 1 - \exp(-\lambda t^\rho).$$

The discrete Weibull distribution (type 1) (Nakagawa and Osaki 1975) is obtained by the probability between two integer values on the suporte of continuous Weibull distribution. Let $Y \in \mathbb{N}$ follow a discrete Weibull distribution. Then its probability mass function is given by

$$\begin{aligned}\Pr(Y = y) &= \int_y^{y+1} f_T(t)dt = F_T(y+1) - F_T(y) \\ &= \exp(-\lambda y^\rho) - \exp(-\lambda(y+1)^\rho) \\ &= q^{y^\rho} - q^{(y+1)^\rho},\end{aligned} \tag{4.5}$$

where $q = \exp(-\lambda) \in (0, 1)$ and $\rho > 0$.

The mean and variance of a discrete Weibull distribution are given by

$$\mathrm{E}(Y) = \sum_{j=1}^{\infty} q^{j^\rho} \quad \text{and} \quad \mathrm{Var}(Y) = 2\sum_{j=1}^{\infty} jq^{j^\rho} - \mathrm{E}(Y) - \mathrm{E}(Y)^2,$$

for which there are no closed expressions (Klakattawi, Vinciotti, and Yu 2018).

This discretization method can be used to discretize any continuous distribution. From the special cases of Weibull, we have correspondents specials cases of the discretized distribution: discrete exponential, when $\rho = 1$ (the popular geometric distribution) and discrete Rayleigh, when $\rho = 2$.

The main disadvantage of the discrete Weibull is that there is no easy interpretation of the model parameters q and ρ. In fact, there is no genuine dispersion and location parameter, the dispersion as well as the expectation is controlled by both q and ρ.

4.2.4 Generalized Poisson distribution

The generalized Poisson is another generalization of the Poisson distribution resulting from the limiting form of the generalized negative binomial distribution (Consul and Jain 1973). Let Y be a random variable according to the generalized Poisson distribution. Then its probability mass function is given by

$$\Pr(Y = y) = \begin{cases} \left[\lambda(\lambda + y\gamma)^{y-1}\exp(-\lambda - y\gamma)\right]/y!, & y = 0, 1, 2, \ldots \\ 0 & y > m, \text{when } \gamma < 0, \end{cases} \tag{4.6}$$

where $\lambda > 0$, $\max(-1, -\lambda/4) \leq \gamma \leq 1$ and m is the largest integer value for which $\lambda + m\gamma > 0$ when γ is negative (Consul and Famoye 1992). The mean and variance is obtained by $\mathrm{E}(Y) = \lambda(1-\gamma)^{-1}$ and $\mathrm{Var}(Y) = \lambda(1-\gamma)^{-3}$. The Poisson distribution is a particular case when $\gamma = 0$.

To obtain the model in mean parametrization, let's assume

$$\lambda = \frac{\mu}{1+\sigma\mu} \quad \text{and} \quad \gamma = \frac{\sigma\mu}{1+\sigma\mu}.$$

The resulting probability mass function can be written as

$$\Pr(y) = \left(\frac{\mu}{1+\sigma\mu}\right)^y \frac{(1+\sigma y)^{y-1}}{y!} \exp\left[-\mu\frac{(1+\sigma y)}{(1+\sigma\mu)}\right], \tag{4.7}$$

where $\sigma > \min(-1/y, -1/\mu)$, when $\sigma < 0$. The mean and variance under this parametrization are given by

$$\mathrm{E}(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = \mu(1+\mu\sigma)^2. \tag{4.8}$$

Thus, this distribution can model under-, when $\min(-1/y, -1/\mu) < \sigma < 0$ and overdispersion, when $\sigma > 0$. The Equation (4.7) reduces to the Poisson distribution when $\sigma = 0$. Regression models based on the generalized Poisson are challenging by the presence of unusual parameter space for underdispersion.

4.2.5 Double Poisson distribution

The double Poisson distribution has been proposed by Efron (1986) based on the double exponential family. Double exponential families allow the introduction of a second parameter that controls variance independently of the mean. Following Efron (1986), a random variable Y double Poisson distributed has the probability mass function given by

$$\Pr(Y = y) = \sqrt{\varphi^{-1}} \exp\left(-\frac{\mu}{\varphi}\right) \left(\frac{\exp(-y)y^y}{y!}\right) \left(\frac{e\mu}{y}\right)^{y/\varphi} \frac{1}{K(\mu, \varphi)}, \tag{4.9}$$

where $\mu > 0$, $\varphi > 0$ and $K(\mu, \varphi)$ is a normalizing constant that can be calculated as

$$K(\mu, \varphi) = \sum_{j=0}^{\infty} \sqrt{\varphi^{-1}} \exp\left(-\frac{\mu}{\varphi}\right) \left(\frac{\exp(-j)j^j}{j!}\right) \left(\frac{e\mu}{j}\right)^{j/\varphi} \approx 1 + \frac{\varphi - 1}{12\mu}\left(1 + \frac{\varphi}{\mu}\right). \tag{4.10}$$

The closed form approximation to the $K(\mu, \varphi)$ (4.10) demostrated by Efron (1986) is reasonable for large values of μ, but is a poor approximation for small values (Zou, Geedipally, and Lord 2013).

The expected value and the variance referring to the double Poisson distribution have no closed form, but Efron (1986) shows that it can be approximated as

$$\text{E}(Y) \approx \mu \quad \text{and} \quad \text{Var}(Y) \approx \varphi\mu.$$

So, this distribution can model under- ($0 < \varphi < 1$), over- ($\varphi > 1$) and equidipersion ($\varphi = 1$, special case Poisson). The disadvantages of using double Poisson to count data analysis are related to the failure to obtain exact results and to the complexity of the normalizing constant (Lindsey 1996, sec. 2.3.3; Winkelmann 2008, sec. 2.6.4).

4.2.6 Poisson-Tweedie distribution

Poisson mixture models (Winkelmann 2008, sec. 2.5; Jørgensen 1997, sec. 4.6.1), also called two-stage models (Hinde and Demétrio 1998) are widely applied to model overdispersed count data. These models are specified hierarchically as $Y \sim \text{Poisson}(\lambda)$ and $\lambda \sim \mathcal{D}(\boldsymbol{\theta})$. The standard example for counts is the negative binomial distribution, where the mean parameter λ is assumed to be gamma distributed.

A general case of Poisson mixture models is a Poisson-Tweedie case (Jørgensen 1997). The Poisson-Tweedie family is given by following the hierarchical specification

$$Y \mid Z \sim \text{Po}(Z) \quad \text{with} \quad Z \sim \text{Tw}_p(\mu, \omega), \tag{4.11}$$

where $\text{Tw}_p(\mu, \omega)$ denotes a Tweedie distribution, a proper exponential dispersion model (Jørgensen 1997) with probability function given by

$$f_Z(z; \mu, \omega, p) = a(z, \omega, p) \exp[(z\psi - k_p(\psi))/\omega].$$

where $\mu = k_p'(\psi)$ is the expectation, $\omega > 0$ is the dispersion parameter, ψ is the canonical parameter and $k_p(\psi)$ is the cumulant function. Furthermore, $\text{var}(Z) = \omega V(\mu)$ where $V(\mu) = k_p''(\psi)$ is the variance function. Tweedie densities are characterized by power variance functions of the form $V(\mu) = \mu^p$, where $p \in (-\infty, 0] \cup [1, \infty)$. The support of the distribution depends on the value of the power parameter. For $p \geq 2$, $1 < p < 2$ and $p = 0$ the support corresponds to the positive, non-negative and real values, respectively.

The probability mass function for the Poisson-Tweedie for $p > 1$ is given by integrating out z, that is

$$\Pr(Y = y) = \int_0^{\infty} \frac{z^y \exp(-z)}{y!} a(z, \omega, p) \exp[(z\psi - k_p(\psi))/\omega] dz. \tag{4.12}$$

The probability distribution (4.12) cannot be written in a closed form, apart from the special case negative binomial distribution ($p = 2$). However, due to hierarchical specification (4.11), the mean and variance can easily be obtained,

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}[\mathrm{E}(Y|Z)] = \mu \\ \mathrm{Var}(Y) &= \mathrm{Var}[\mathrm{E}(Y|Z)] + \mathrm{E}[\mathrm{Var}(Y|Z)] = \mu + \omega\mu^p. \end{aligned} \tag{4.13}$$

Besides the negative binomial ($p = 2$), special cases include Hermite ($p = 0$), Neymann type-A ($p = 1$), Pólya-Aeppli ($p = 1,5$), Poisson compound Poisson ($1 < p < 2$) and Poisson-inverse Gaussian ($p = 3$) (Bonat et al. 2018; Bonat, Zeviani, and Ribeiro Jr 2017).

The Poisson-Tweedie model only handle overdispersion ($\omega > 0$). However, Bonat et al. (2018) noted that variance can be smaller than the expectation allowing $\omega < 0$. So, they proposed the extended Poisson-Tweedie model, specified using only second-order moments assumptions. The extended Poisson-Tweedie model can deal with over- ($\omega > 0$), equi- ($\omega = 0$) and underdispersion ($\omega < 0$). The only restriction to have a proper model is that $\mathrm{Var}(Y) > 0$ implying that $\omega > -\mu^{(1-p)}$.

4.3 Comparing count distributions

In order to explore and compare the flexibility of the models aforementioned to deal with real count data, we compute indexes for dispersion (DI), zero-inflation (ZI) and heavy-tail (HT), which are respectively given by

$$\mathrm{DI} = \frac{\mathrm{Var}(Y)}{\mathrm{E}(Y)}, \quad \mathrm{ZI} = 1 + \frac{\log \Pr(Y=0)}{\mathrm{E}(Y)} \quad \text{and} \quad \mathrm{HT} = \frac{\Pr(Y = y+1)}{\Pr(Y=y)} \quad \text{for} \quad y \to \infty.$$

As discussed in Chapter 3, these indexes are defined in relation to the Poisson distribution. Thus, the dispersion index indicates overdispersion for $\mathrm{DI} > 1$, underdispersion for $\mathrm{DI} < 1$ and equidispersion for $\mathrm{DI} = 1$; the zero-inflation index indicates zero-inflation for $\mathrm{ZI} > 0$, zero-deflation for $\mathrm{ZI} < 0$ and no excess of zeros for $\mathrm{ZI} = 0$; and the heavy-tail index indicates a heavy-tail distribution for $\mathrm{HT} \to 1$ when $y \to \infty$. These indices have been widely applied to explore distributions. Bonat et al. (2018) used them to study the Poisson-Tweedie family whereas Ribeiro Jr et al. (2018) and Luyts et al. (2018) used them to explore the reparametrized COM-Poisson distribution and the discrete-Weibull distribution, respectively. A brief discussion of these indexes can be found in Puig and Valero (2006).

To study and compare the flexibility of the distributions, we considered all two-parameter distributions described in Section 4.2 and three special cases of three-parameter Poisson-Tweedie distribution: Poisson compound Poisson ($p = 1.1$), negative binomial ($p = 2$) and Poisson inverse-Gaussian ($p = 3$). The dispersion parameters of the count distributions were set to have $\mathrm{DI} = 0.2, 0.5, 1, 2, 5$ and 10 when the means equals 10 to allow comparisons of distribution. Table 4.1 shows the dispersion parameters found for each scenario. For the CMP and DWe distribution, we could not find the parameter values to have dispersion index equal to 10 when the expected value is 10. For the location parameters, we consider a sequence to have the expected values between 0 and 50. The expected values and the variances for COM-Poisson (CMP), Gamma-count (GCT), discrete-Weibull (DWe) and double Poisson (DPo) distributions,

Table 4.1. Dispersion Scenarios: dispersion parameters of CMP, GCT, DWe, GCT, GPo, DPo, and PTw$_p$ distributions when $\mathrm{E}(Y)$ is set to 10 and $\mathrm{DI}(Y) = 0.2, 0.5, 1, 2, 5,$ and 10.

	Scenario 1 (DI = 0.2)	Scenario 2 (DI = 0.5)	Scenario 3 (DI = 1)	Scenario 4 (DI = 2)	Scenario 5 (DI = 5)	Scenario 6 (DI = 10)
CMP(ν)	5.2059	2.0526	1.0000	0.4687	0.1292	–
GCT(α)	5.4221	2.0785	1.0000	0.4643	0.1370	0.0213
DWe(ρ)	9.0735	5.4642	3.7150	2.5189	1.5145	–
GPo(σ)	-0.0553	-0.0293	0.0000	0.0414	0.1236	0.2162
DPo(φ)	0.2001	0.5002	1.0000	1.9856	5.1625	13.9295
PTw$_{1.1}$(ω)	–	–	–	0.7943	3.1773	7.1490
PTw$_{2.0}$(ω)	–	–	–	0.1000	0.4000	0.9000
PTw$_{3.0}$(ω)	–	–	–	0.0100	0.0400	0.0900

are computed numerically using

$$\mathrm{E}(Y) = \sum_{y=0}^{500} y \Pr(Y = y) \quad \text{and} \quad \mathrm{Var}(Y) = \sum_{y=0}^{500} \{[y - \mathrm{E}(Y)]^2\} \Pr(Y = y).$$

For the generalized Poisson (GPo) and Poisson-Tweedie (PTw) distributions, the expected values and the variances have closed forms (Equations (4.8) and (4.13)). To obtain the HT and ZI for the Poisson-Tweedie distributions, we approximate the value of integral (4.12) using the Gauss-Laguerre method with 100 nodes, following the codes provide by Bonat et al. (2018). Since the Poisson-Tweedie cannot deal with underdispersion, in these scenarios we present only the other distributions. Figures 4.1, 4.2 and 4.3, display the indexes by the different scenarios.

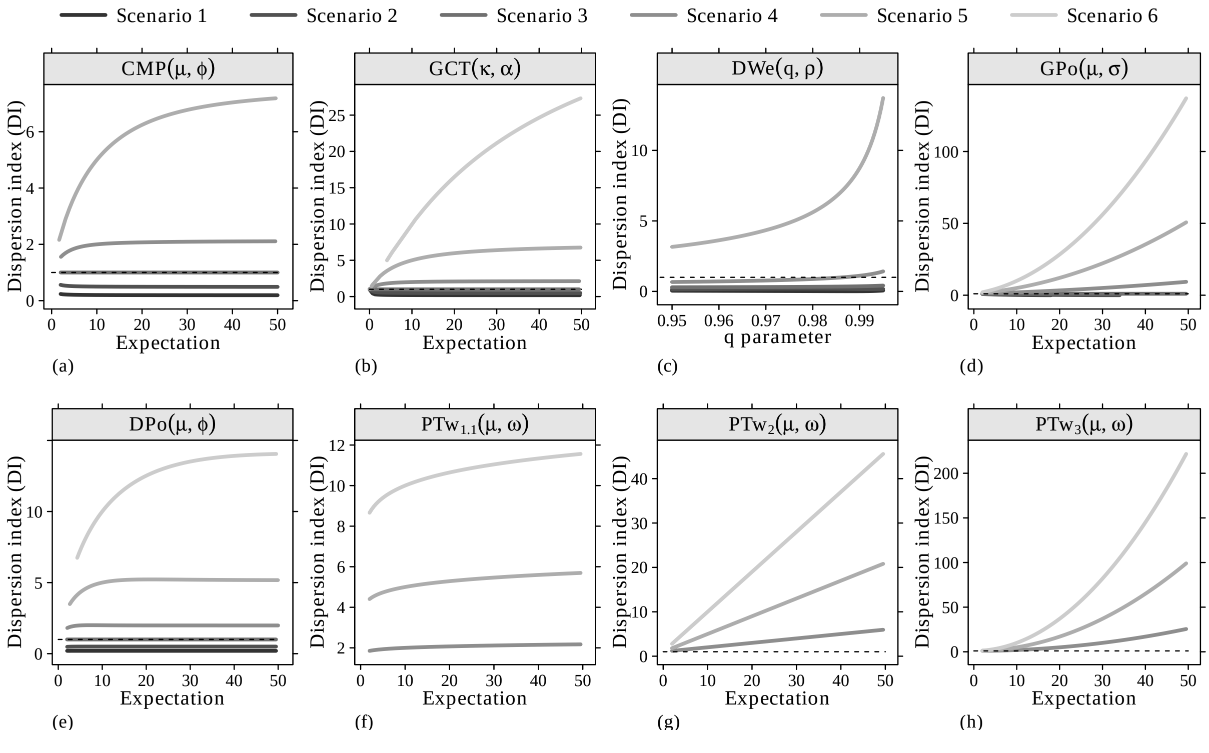

Figure 4.1. Dispersion index as function of expected values by different scenarios and count distributions. Dotted lines (DI = 1) represent the Poisson distribution.

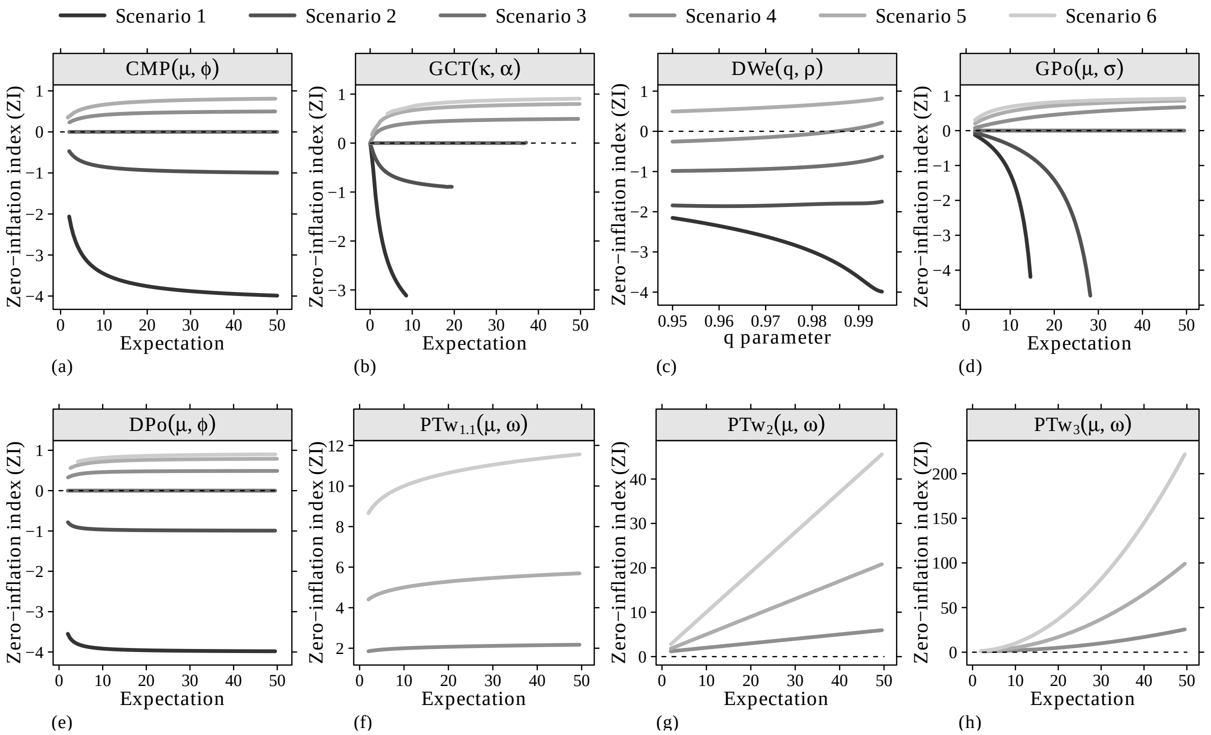

Figure 4.2. Zero-inflation index as function of expected values by different scenarios and count distributions. Dotted lines (ZI = 0) represent the Poisson distribution.

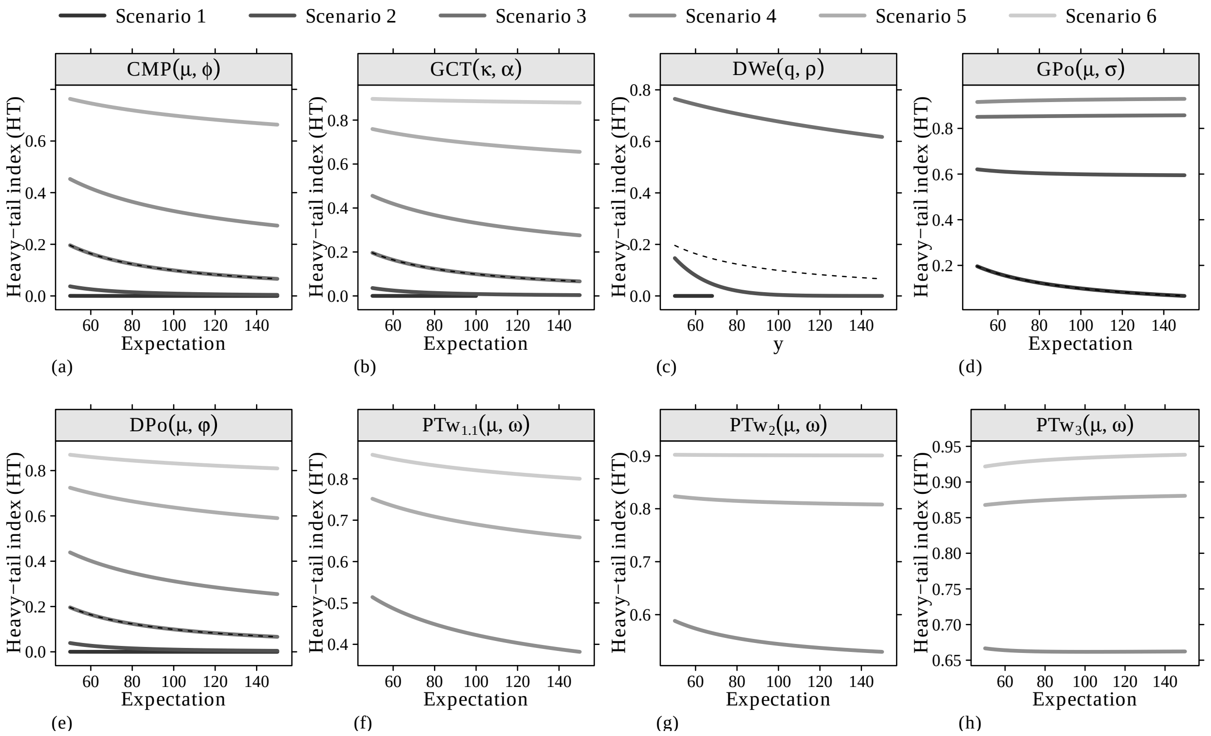

Figure 4.3. Heavy-tail index for some extreme values of the random variable Y considering $\mathrm{E}(Y) = 10$ by different scenarios and count distributions. Dotted lines ($\mathrm{HT} = \mu(y+1)^{-1}$) represent the Poisson distribution.

The DIs in Figure 4.1 show that the CMP, GCT, DPo, and $\text{PTw}_{1.1}$ (only for overdispersion) are quite similar. In general, for these distributions, the indexes depend slightly on the expected values and tend to stabilize for large expected values. Consequently, the mean-variance relationship is proportional to the dispersion parameter value. In fact, if we look at the (approximate) variances for these distributions, they follow the same structure. The DIs for PTw_3 and GPo distributions shows that distributions are suitable to handle very strong overdispersion. For the DWe distribution, the x-axis is the q parameter, so the interpretation is not straightforward.

The ZIs in Figure 4.2 show that CMP, GCT, DWe, GPo, and DPo can handle a limited amount of zero-inflation, in cases of overdispersion ($\text{DI} < 1$) and it is suitable for zero-deflation, in cases of underdispersion ($\text{DI} > 1$). For the GCT distribution on the underdispersion scenarios ($\text{DI} = 0.5$ and $\text{DI} = 0.5$), the probability at zero tends to 0 as the mean increases, hence the ZI $\rightarrow -\infty$. That is why the curves are dropped for the GCT distribution. Also in the underdispersion scenarios, the shape of the curves for the GPo distribution is interesting. Analytically, the ZI for GPo distribution is $(-1 - \sigma\mu)^{-1}$, which has a vertical asymptote at $\mu = -1/\sigma$, $\lim_{\mu \to -1/\sigma} \text{ZI}(\mu)$ is equals to $-\infty$ from the left and ∞ from the right. However, there is no distribution when $\mu > -1/\sigma$ (see Equation (4.7)). The ZI's for PTw distributions increase quickly as the mean increases, indicating that distributions are suitable to deal with zero-inflated count data.

Finally, HTs in Figure 4.3 indicates that CMP, GCT, DWe, DPo and $\text{PTw}_{1.1}$ distributions are in general light-tailed distributions i.e. HT $\rightarrow 0$ for $y \rightarrow \infty$. Actually, the HT for the DWe distributions tends to 1 for small values of the expectation (Luyts et al. 2018). The indexes for GPo and PTw_p ($p = 1, 2$) distributions increases with increasing mean, showing that the model is especially suitable to deal with heavy-tailed count data.

4.4 Regression models and estimation

Suppose y_i, $i = 1, \dots, n$ is a set of independents realizations of Y_i according to distribution $\mathcal{D}(.,.)$ and $\boldsymbol{x}_i$ a vector of known covariates. The regression models based on the distributions aformentioned is defined by

$$Y_i \sim \mathcal{D}(g^{-1}(\eta_i), \theta), \quad \text{with} \quad \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

where θ is a general notation for the dispersion (or second) parameter of the distribution $\mathcal{D}$ and $g(.)$ is a suitable link function.

Table 4.2 summarizes the main properties of the count distributions considered, presents the link functions adopted, and establishes the notation. Similar results (exact, approximated or limiting) can be directly linked to the results presented in Figure 4.1.

Except for the Poisson-Tweedie, the parameter estimation is based on the maximum likelihood method and the inference is done using the standard machinery of likelihood inference, including likelihood ratio tests for model comparison and Wald-tests for testing individual (or groups of) parameters. Since the derivatives of the log-likelihood function cannot be obtained analytically for any of these models, we compute them by central finite differences using the Richardson method as implemented in package `numDeriv` (Gilbert and Varadhan 2016) in `R` (R Core Team 2018). The normalizing constants present in the COM-Poisson and double

Table 4.2. Summary of the considered regression models for analysis of dispersed count data.

	COM-Poisson	Gamma-Count	Discrete Weibull
Notation	$\text{CMP}(\mu_i, \nu)$	$\text{GCT}(\kappa_i, \alpha)$	$\text{DWe}(q_i, \rho)$
Linear predictor	$\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$	$\log(\kappa_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$	$\log(-\log(q_i)) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$
Dispersion parameter	$\log(\nu)$; $\nu > 0$	$\log(\alpha)$; $\alpha > 0$	$\log(\rho)$; $\rho > 0$
Expectation	$\approx \mu_i$	$\overset{a}{\approx} \kappa_i$	—
Variance	$\approx \mu_i/\nu$	$\overset{a}{\approx} \kappa_i/\alpha$	—
Dispersion index (DI)	$\approx 1/\nu$	$\overset{a}{\approx} 1/\alpha$	—

	Generalized Poisson	Double Poisson	Poisson-Tweedie
Notation	$\text{GPo}(\mu_i, \sigma)$	$\text{DPo}(\mu_i, \varphi)$	$\text{PTw}_p(\mu_i, \omega)$
Linear predictor	$\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$	$\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$	$\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$
Dispersion parameter	σ; $\sigma > c^*$	$\log(\varphi)$; $\varphi > 0$	ω $\omega > 0$
Expectation	μ_i	$\approx \mu_i$	μ_i
Variance	$\mu_i(1 + \sigma\mu_i)^2$	$\approx \varphi\mu_i$	$\mu_i(1 + \omega\mu_i^{p-1})$
Dispersion index (DI)	$(1 + \sigma\mu_i)^2$	$\approx \varphi$	$1 + \omega\mu_i^{p-1}$

$c^* = \min[-\max(y_i^{-1}), -\max(\mu_i^{-1})]$; $\overset{a}{\approx}$ asymptotically when $T \to \infty$.

Poisson model are calculated by truncating the series (i.e., the sum of the first k terms). The computational routines for fitting these models are implemented in `R`, using `C++` to compute the normalizing constant for CMP and DPo, and organized in a `R` package (`flexcm`)[1].

For the Poisson-Tweedie, the non-trivial restriction on the power parameter space and the presence of an intractable integral in the probability mass function makes the estimation by maximum likelihood method hard to use. Furthermore, by using the maximum likelihood method we can model only overdispersio. Therefore, we use only the second-order moment assumptions for this class, i.e. $\text{E}(Y_i) = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \omega\mu_i^p$. Note that this specification allows us to handle under- and overdispersion and eliminate the non-trivial constraint in the power parameter space. The only restriction to have a proper model is that $\omega > -\mu_i^{1-p}$. This approach is called by extended Poisson-Tweedie (Bonat et al. 2018).

Estimation and inference for the extended Poisson-Tweedie model are based on the estimating function approach. Namely, quasi-score functions are used to estimate the regression parameters $\boldsymbol{\beta}$ and Pearson estimating function are used to estimate the variance parameters ω and p. This estimation method is implemented in the `mcglm` package (Bonat 2018) as a special case of the multivariate covariance generalized linear models (Bonat and Jørgensen 2016).

[1] Available on GitHub https://github.com/jreduardo/flexcm

4.5 Data analyses

In this section, we analyzed two data sets to illustrate and compare the applications of the flexible regression models. The data sets and `R` code for their analysis are available in the appendix.

4.5.1 *Sitophilus zeamaus* experiment

As the first example, we analyse the *Sitophilus zeamaus* data set introduced in Section 2.5. This dataset results from a completely randomized experiment with ten replicates and four treatments. To analyze the number of insects (Y_{ij}), we consider the linear predictor $\eta_{ij} = \beta_0 + \tau_j$, where $i = 1, 2, \ldots, 10$ and j refers to the treatments (control, leaf, branch, and seed).

The parameter estimates, associated standard errors and goodness-of-fit measures (maximized log-likelihood and Akaike information criterion) are given in Table 4.3. Note that for the PTw model, we could calculate the log-likelihood function since the estimated parameters ($\hat{p} = 1.4031$ and $\hat{\omega} = 0.3476$) correspond to an existing Poisson-Tweedie distribution. Furthermore, except for DWe model, all models have large improvements related to the standard Poisson model, whose maximum log-likelihood is equal to -130.5828. The likelihood ratio tests for $\text{H}_0 : \log(\theta) = 0$ (where θ is the associated dispersion parameter) for the models nesting the Poisson model (i.e. CMP, GCT, GPo, DPo and PTw) indicates strong evidence to reject the null hypothesis.

In terms of practical inference, all models lead to the same conclusion – the extract prepared with seeds of *Annano mucosa* drastically decrease the *Sitophilus zeamais* progeny while the other solutions (control, leaves and branch) had the same effect (see regression coefficients in Table 4.3). The same conclusion is reached by Demétrio, Hinde, and Moral (2014) using a quasi-Poisson specification (Wedderburn 1974).

Table 4.3 also shows the disadvantage of the DWe model in terms of interpretation. Whereas the others models are parameterized in terms of mean, even approximately or asymptotically, the DWe model parameters are on a not easily interpretable scale.

Fitted counts with 95% confidence intervals for the six models are given in Figure 4.4. The results are practically identical for CMP, GCT, GPo, DPo and PTw. Only for DWe the results are slightly lower than the values reported for the others. Thus, along with the log-likelihoods presented in Table 4.3, it seems that the DWe model is not suitable to fit this dataset.

Figure 4.4(b) shows the mean-variance relationship for the six fitted models. To obtain these curves, we computed the expected values and the the variances for the location parameter varying from $g^{-1}(\eta_{\text{lwr}})$ to $g^{-1}(\eta_{\text{upr}})$ (lower and upper of the linear predictor confidence interval), and dispersion parameter fixed at the maximum likelihood estimate. The results show that the variance increases linearly for CMP, GCT and DPo and polynomially for DWe, GPo and PTw. In fact, the polynomials for GPo and PTw are $f(x) = x(1 + x/50)^2$ and $f(x) = x + 7x^{1.03}/20$, respectively.

Table 4.3. *Sitophilus zeamais* data: Parameter estimates (Est) and standard errors (SEs) for the six fitted count models.

		Estimates (Standard Errors)		
Parameter		COM-Poisson	Gamma-Count	Discrete Weibull
Dispersion				
	$\log(\nu)$	−0.9272 (0.2628)[a]	–	–
	$\log(\alpha)$	–	−0.9273 (0.2635)[a]	–
	$\log(\rho)$	–	–	1.0348 (0.1361)[a]
Regression				
	β_0	3.4497 (0.0885)[a]	3.4255 (0.0911)[a]	−9.8805 (1.4027)[a]
	τ_{leaf}	−0.0064 (0.1254)	−0.0066 (0.1281)	−0.0487 (0.4478)
	τ_{branch}	−0.0522 (0.1268)	−0.0535 (0.1296)	0.0717 (0.4474)
	τ_{seed}	−3.2552 (0.2801)[a]	−4.0157 (0.6629)[a]	7.6069 (1.0371)[a]
LogLik		−121.6334	−121.6509	−128.7893
AIC		253.2668	253.3017	267.5787

		Estimates (Standard Errors)		
Paramater		Generalized Poisson	Double Poisson	Poisson-Tweedie
Power				
	p	–	–	1.4031 (0.5703)[a]
Dispersion				
	α	0.0194 (0.0070)[a]	–	–
	$\log(\varphi)$	–	0.8659 (0.2444)[a]	–
	ω	–	–	0.3476 (0.6699)
Regression				
	β_0	3.4500 (0.0908)[a]	3.4503 (0.0868)[a]	3.4500 (0.0872)[a]
	τ_{leaf}	−0.0064 (0.1285)	−0.0064 (0.1230)	−0.0064 (0.1235)
	τ_{branch}	−0.0521 (0.1289)	−0.0521 (0.1244)	−0.0521 (0.1246)
	τ_{seed}	−3.3547 (0.3211)[a]	−3.5644 (0.5611)[a]	−3.3547 (0.3624)[a]
LogLik		−122.2840	−121.7930	−121.8466
AIC		254.5680	253.5860	255.6932

Est (SE)[a] indicates $|\text{Est/SE}| > 1,96$.

4.5.2 Bromeliad experiment

The second example relates to the bromeliad experiment of Section 2.5. The number of leaves per experimental unit Y_{ijk} was recorded for different treatments (Xaxim and alternative substrates) at six dates after planting. To analyze this data set, we dropped the data from the first date (4 days after planting) in order to avoid the nonlinearity (sigmoidal) relationship between response and time (see Figure 2.4). In addition, the practical interest relies on the behavior of the plants close to the time when it is suitable for commercialization, so there is no relevant information at the 4 days after planting.

To model Y_{ijk}, we assume it distributed according to the six flexible distributions and consider the following linear predictors

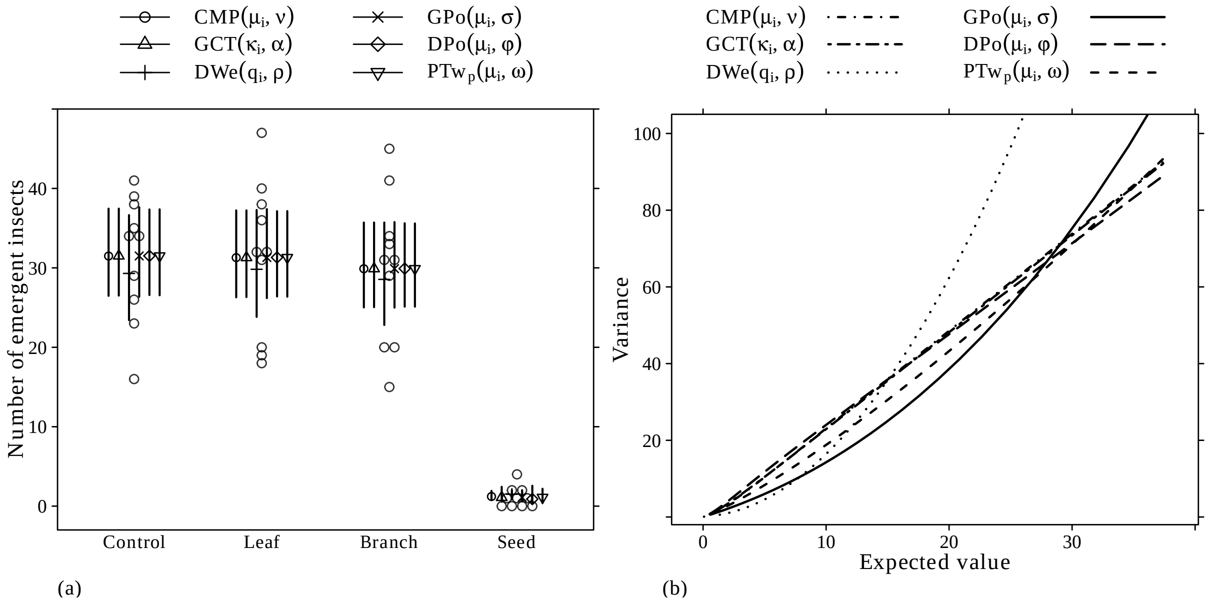

Figure 4.4. (a) Scatterplots of the observed data and fitted values with 95% confidence intervals and (b) Fitted mean and variance relationship for the six models.

$$
\begin{aligned}
&\text{Varying } \beta_0: && \eta_{ijk} = \beta_0 + \gamma_i + \tau_j + \beta_1 \mathtt{x}_{1k} + \beta_2 \mathtt{x}_{2k} \\
&\text{Varying } \beta_0 \text{ and } \beta_1: && \eta_{ijk} = \beta_0 + \gamma_i + \tau_j + (\beta_1 + \delta_{1j})\mathtt{x}_{1k} + \beta_2 \mathtt{x}_{2k} \\
&\text{Varying } \beta_0,\ \beta_1 \text{ and } \beta_2: && \eta_{ijk} = \beta_0 + \gamma_i + \tau_j + (\beta_1 + \delta_{1j})\mathtt{x}_{1k} + (\beta_2 + \delta_{2j})\mathtt{x}_{2k},
\end{aligned}
$$

where i, j, and k refers to block, treatment and time, respectively. The covariates $\mathtt{x}_{1k}$ and $\mathtt{x}_{2k}$ are the orthogonal linear and quadratic polynomials of log(time).

Descriptive analysis of the number of leaves in the bromeliad experiment in Section 2.5 suggested strong underdispersion. Thus, we consider only first and second moments for the Poisson-Tweedie model. The estimation method implemented in `mcglm` package crashes to fit this dataset when considering free power parameter. However, to fixed power parameter between 0.93 and 1.1 we are able to estimate ω and $\boldsymbol{\beta}$. In Figure 4.5, we show the profile pseudo-log-likelihood for the power parameter p, considering the three linear predictors. The results suggest that a special case Neyman Type A (NTA) ($p = 1$) could be a good choice to fit this dataset.

The maximum likelihood achieved by the full parametric model is given in Table 4.4. The minus twice log-likelihood obtained from the Poisson model was equal to 439.4984 with 82 degrees of fredoom and the deviance was equal to 1.82. Here, there is clearly overwhelming evidence against the Poisson assumption. The CMP, GCT and DPo models fit the data better than the DWe and GPo models in terms of likelihood. The latter model gives a poor fit for the underdispersion. In practical terms, this leads to different conclusions about the hypothesis H_0: $\beta_{1j} = 0$. For CMP, GCT, DWe and DPo models there is evidence to reject this hypothesis whereas for GPo there is no evidence to reject H_0.

The estimated parameters and associated standard errors for the full parametric models and for the NTA model (Poisson-Tweedie with fixed power parameter at 1) are presented in Table 4.5. All models indicate strong underdispersion confirming the findings in the descriptive analysis. Therefore, since the CMP, GCT, and DPo models are approximately indistinguishable in terms of moments, it is interesting to note that $\hat{\nu} \approx 1/\hat{\varphi} \approx \hat{\alpha}$.

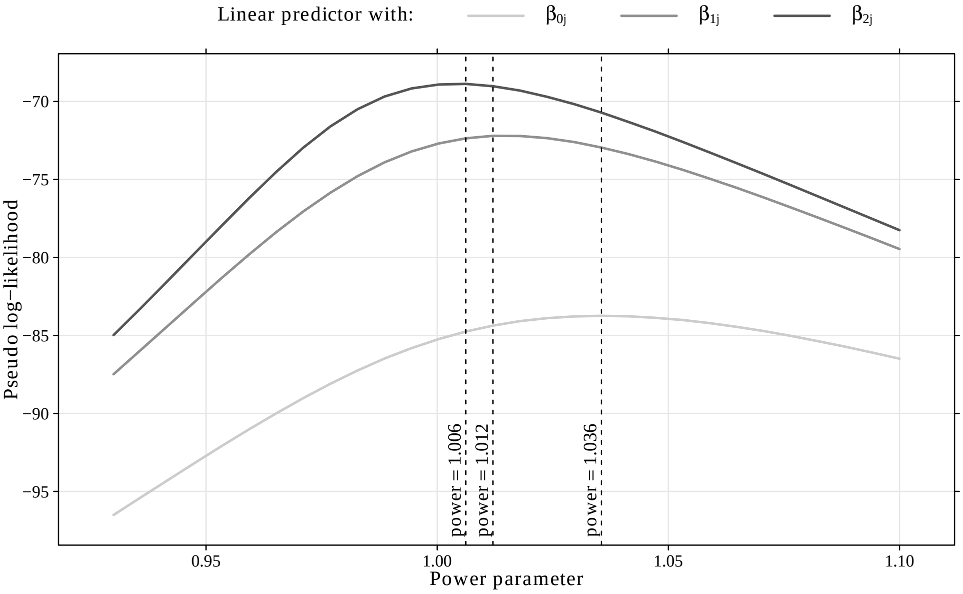

Figure 4.5. Profile pseudo-likelihood for the power parameter of the extended Poisson-Tweedie model with different linear predictors.

Table 4.4. Bromelia data: likelihoods (−2×logLik) and number of parameters (#p) for the three predictors and five full parametric flexible models fitted to the data set.

	Minus twice log-likelihood (#p)				
Term varying	CMP	GCT	DWe	GPo	DPo
Intercept (β_{0j})	169.50 (11)[a]	168.16 (11)[a]	178.03 (11)[a]	216.90 (11)[a]	169.71 (11)[a]
Slope (β_{1j})	143.93 (15)[a]	143.15 (15)[a]	150.53 (15)[a]	211.34 (15)	144.03 (15)[a]
Quadratic (β_{2j})	135.97 (19)	136.30 (19)	144.57 (19)	209.98 (19)	135.98 (19)

[a] indicates the coefficients are significantly different from 0 in the likelihood ratio test at 5% level.

The regression coefficients are quite similar among the CMP, GCT, GPo, DPo and NTA. The standard errors obtained from the GPO model are larger than the others, indicating that this model does not fit well to the underdispersion of this dataset. Therefore, except for the GPo models, the individual Wald tests indicate that intercept, linear and quadratic terms are different for each substrate.

Figure 4.6(a) presents the fitted curves with 95% confidence bands for the six modeling strategies and Figure 4.6(b) the fitted curves for each alternative substrate considering the CMP model. The results support that the growth rate of the expected number of leaves is higher for the bromeliads grown with the substrate composite with *Pinus* bark, being this the preferred substrate to replace the composite with Xaxim.

In Figure 4.7, we show the differences between the upper and lower limits of 95% confidence bands for all considered models in order to compare them. The confidence bands are, in general, larger for GPo model. For CMP, GCT, DWe, DPo, and NTA, the confidence intervals are quite similar, with small differences for DWe. The NTA gives slightly larger intervals than the others, this is because we made only mean-variance assumptions to fit this model.

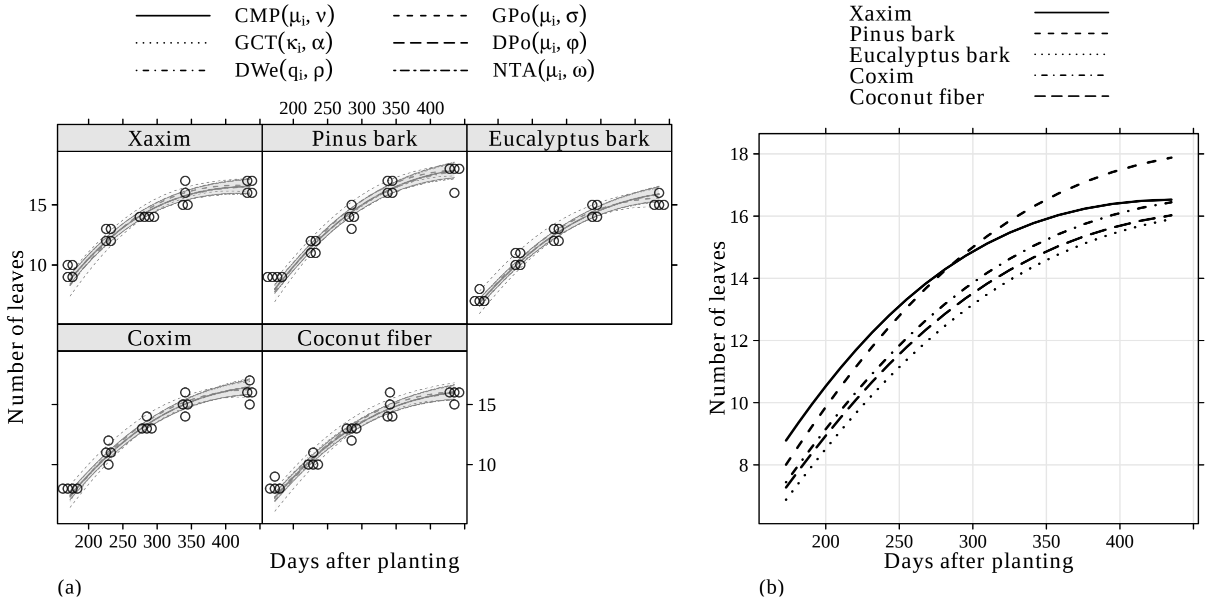

Figure 4.6. (a) Scatterplots of the observed data and fitted values with 95% confidence bands and (b) Fitted curves for each treatment considering the CMP model.

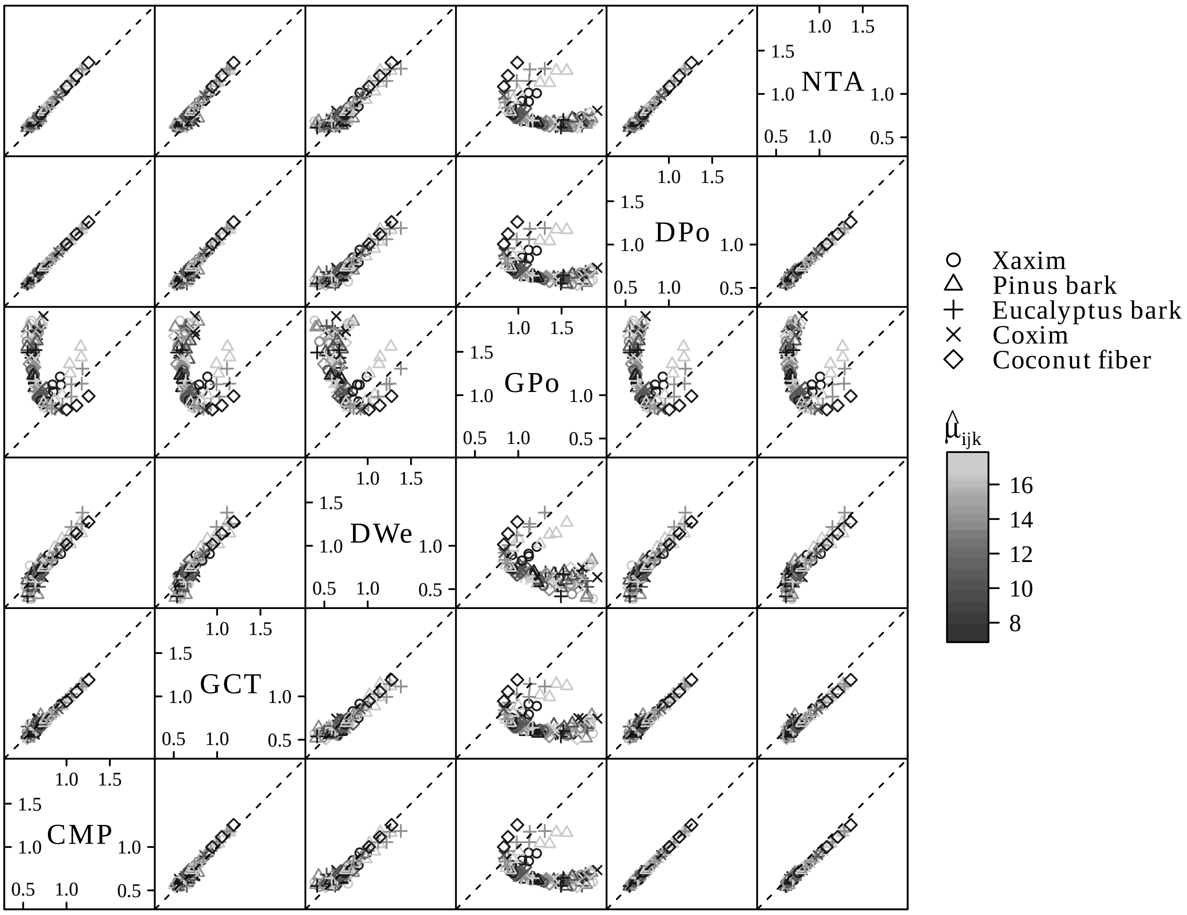

Figure 4.7. Pairwise plot of the differences between the upper and lower limits of 95% confidence bands for all considered models. Symbols refer to the treatments and colors refer to the fitted values.

Table 4.5. Bromelia data: Parameter estimates (Est) and standard errors (SEs) for the six fitted count models.

Paramater	COM-Poisson	Gamma-Count	Discrete Weibull
Dispersion			
$\log(\nu)$	3.9689 (0.1430)$^{\mathrm{a}}$	–	–
$\log(\alpha)$	–	4.4197 (0.2097)$^{\mathrm{a}}$	
$\log(\rho)$	–	–	3.5722 (0.0885)$^{\mathrm{a}}$
Regression			
β_0	2.5865 (0.0110)$^{\mathrm{a}}$	2.6224 (0.0102)$^{\mathrm{a}}$	−93.9304 (8.3372)$^{\mathrm{a}}$
γ_{II}	0.0217 (0.0110)$^{\mathrm{a}}$	0.0233 (0.0103)$^{\mathrm{a}}$	−0.8595 (0.4016)$^{\mathrm{a}}$
γ_{III}	−0.0285 (0.0113)$^{\mathrm{a}}$	−0.0271 (0.0106)$^{\mathrm{a}}$	1.1354 (0.4084)$^{\mathrm{a}}$
γ_{IV}	0.0400 (0.0111)$^{\mathrm{a}}$	0.0397 (0.0103)$^{\mathrm{a}}$	−1.2435 (0.4140)$^{\mathrm{a}}$
τ_{Pinus}	−0.0070 (0.0122)	−0.0068 (0.0115)	0.2847 (0.4283)
$\tau_{\mathrm{Eucaliptos}}$	−0.1415 (0.0125)$^{\mathrm{a}}$	−0.1347 (0.0119)$^{\mathrm{a}}$	4.7957 (0.5927)$^{\mathrm{a}}$
τ_{Coxim}	−0.0855 (0.0127)$^{\mathrm{a}}$	−0.0818 (0.0120)$^{\mathrm{a}}$	2.9691 (0.5097)$^{\mathrm{a}}$
τ_{Coconut}	−0.1099 (0.0128)$^{\mathrm{a}}$	−0.1047 (0.0119)$^{\mathrm{a}}$	3.5371 (0.5007)$^{\mathrm{a}}$
β_1	1.9182 (0.0881)$^{\mathrm{a}}$	1.8552 (0.0838)$^{\mathrm{a}}$	−64.9765 (6.6555)$^{\mathrm{a}}$
δ_{Pinus}	0.5193 (0.1252)$^{\mathrm{a}}$	0.4987 (0.1194)$^{\mathrm{a}}$	−17.7658 (4.6140)$^{\mathrm{a}}$
$\delta_{\mathrm{Eucaliptos}}$	0.6251 (0.1268)$^{\mathrm{a}}$	0.5704 (0.1223)$^{\mathrm{a}}$	−21.3382 (4.9066)$^{\mathrm{a}}$
δ_{Coxim}	0.4874 (0.1305)$^{\mathrm{a}}$	0.4492 (0.1243)$^{\mathrm{a}}$	−18.1868 (4.8095)$^{\mathrm{a}}$
$\delta_{\mathrm{Coconut}}$	0.4789 (0.1319)$^{\mathrm{a}}$	0.4526 (0.1234)$^{\mathrm{a}}$	−14.8219 (4.4322)$^{\mathrm{a}}$
β_2	−0.5064 (0.0403)$^{\mathrm{a}}$	−0.4861 (0.0384)$^{\mathrm{a}}$	17.2198 (2.1367)$^{\mathrm{a}}$

Paramater	Generalized Poisson	Double Poisson	Neymann type-A^1
Dispersion			
σ	−0.0540 (0.0007)$^{\mathrm{a}}$	–	–
$\log(\varphi)$	–	−3.9280 (0.1430)$^{\mathrm{a}}$	–
ω	–	–	−0.9769 (0.0040)$^{\mathrm{a}}$
Regression			
β_0	2.5831 (0.0207)$^{\mathrm{a}}$	2.5864 (0.0110)$^{\mathrm{a}}$	2.5865 (0.0119)$^{\mathrm{a}}$
γ_{II}	0.0239 (0.0135)	0.0215 (0.0110)	0.0217 (0.0120)
γ_{III}	−0.0465 (0.0171)$^{\mathrm{a}}$	−0.0286 (0.0113)$^{\mathrm{a}}$	−0.0286 (0.0121)$^{\mathrm{a}}$
γ_{IV}	0.0276 (0.0144)	0.0399 (0.0111)$^{\mathrm{a}}$	0.0399 (0.0119)$^{\mathrm{a}}$
τ_{Pinus}	0.0053 (0.0235)	−0.0070 (0.0122)	−0.0068 (0.0133)
$\tau_{\mathrm{Eucaliptos}}$	−0.1302 (0.0304)$^{\mathrm{a}}$	−0.1417 (0.0125)$^{\mathrm{a}}$	−0.1420 (0.0138)$^{\mathrm{a}}$
τ_{Coxim}	−0.0760 (0.0276)$^{\mathrm{a}}$	−0.0856 (0.0127)$^{\mathrm{a}}$	−0.0861 (0.0136)$^{\mathrm{a}}$
τ_{Coconut}	−0.1047 (0.0290)$^{\mathrm{a}}$	−0.1099 (0.0128)$^{\mathrm{a}}$	−0.1096 (0.0136)$^{\mathrm{a}}$
β_1	2.1015 (0.1865)$^{\mathrm{a}}$	1.9188 (0.0880)$^{\mathrm{a}}$	1.9185 (0.0964)$^{\mathrm{a}}$
δ_{Pinus}	0.4217 (0.1963)$^{\mathrm{a}}$	0.5188 (0.1252)$^{\mathrm{a}}$	0.5141 (0.1363)$^{\mathrm{a}}$
$\delta_{\mathrm{Eucaliptos}}$	0.3998 (0.2692)	0.6279 (0.1265)$^{\mathrm{a}}$	0.6296 (0.1413)$^{\mathrm{a}}$
δ_{Coxim}	0.3227 (0.2425)	0.4888 (0.1307)$^{\mathrm{a}}$	0.4932 (0.1391)$^{\mathrm{a}}$
$\delta_{\mathrm{Coconut}}$	0.4313 (0.2539)	0.4783 (0.1320)$^{\mathrm{a}}$	0.4713 (0.1400)$^{\mathrm{a}}$
β_2	−0.6186 (0.0853)$^{\mathrm{a}}$	−0.5063 (0.0403)$^{\mathrm{a}}$	−0.5053 (0.0435)$^{\mathrm{a}}$

Est (SE)$^{\mathrm{a}}$ indicates $|\text{Est/SE}| > 1,96$; ^{1}second-order moments assumptions (extended Poisson-Tweedie) with power parameter fixed at 1.

4.6 Discussion

In this chapter, we presented six model strategies for analysis of dispersed count data and compare them using characteristic indexes and applications to the analysis of two experimental data. The genesis of generating probability distributions and their main properties is presented. Although some results are obtained only approximately or asymptotically, it is noted that the COM-Poisson, Gamma-count, and double Poisson models are practically indistinguishable in terms of mean-variance relationship. Moreover, the study of zero-inflation and heavy-tail indexes confirm the similarity. The COM-Poisson and double Poisson models are weighted-type distributions, so the computation of probabilities requires the computation of a normalization constant that makes the time-to-fit slow for big datasets. On the other hand, the Gamma-count distribution has a simple form for a probability function but the interpretation remains on the scale of waiting times. The parametrization of discrete Weibull distribution complicates the study of its flexibility; it seems that it is a distribution useful for small counts. In the two applications, the discrete Weibull model did not fit well compared to the competitive models. The generalized Poisson model, although handle underdispersion, its unusual parametric space complicates the fitting in these cases. However, it was the model that took less time to fit in both applications. Finally, the extended Poisson-Tweedie approach is fast to fit and very flexible. However, it is based only on second-order moments assumption and there is no distribution for underdispersion what makes impossible to compute probabilities, for example. Moreover, when analyzing equidispersion data there is no information about the power parameter which makes its estimation difficult. In such cases, the inference about the regression parameters is independent of the choice of power parameter.

Regarding the analysis of experimental data, for the first experiment, it is clear that the extract prepared with seeds is different from the others and the other solution has no difference even without fit any model. On the other hand, the second application presents several statistical challenges. This is a longitudinal experiment – the same plot was evaluated at different times with a clear evidence of strong underdispersion and nonlinear response over time. To analyze this data, we selected the data only from the region of practical interest and for this region, a polynomial approximation was a good choice. To address the correlated counts, we try to fit mixed models, however, although the observations are from the same individual, there is no evidence of correlation between them, the predicted values for the random effects were very close to zero.

For further research, we are working to allow nonlinear predictors for the mean parameter of these models. There are many experiments in which there is a biological process leading to nonlinear models, and the interpretation of the parameters becomes very useful for practitioners. Furthermore, the original parametrization of discrete Weibull does not seems good to modeling, thus investigating new parametrizations can be useful. Finally, allow the dispersion parameter to be modeled depending on covariates is a natural extension.

References

Bonat, W. H. (2018). “Multiple Response Variables Regression Models in R: The mcglm Package”. In: *Journal of Statistical Software* 84.4, pp. 1–30.

Bonat, W. H. and B. Jørgensen (2016). “Multivariate covariance generalized linear models”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics).*

Bonat, W. H., B. Jørgensen, C. C. Kokonendji, and J. Hinde (2018). “Extended Poisson-Tweedie: properties and regression model for count data”. In: *Statistical Modelling* 18.1, pp. 24–49.

Bonat, W. H., W. M. Zeviani, and E. E. Ribeiro Jr (2017). *Regression Models for Count Data: beyond Poisson model.* Goiás, Brazil: XV EMR - Brazilian Regression Model School.

Cameron, A. C. and P. K. Trivedi (2013). *Regression Analysis of Count Data.* 2nd edition. Econometric Society Monographs. New York: Cambridge University press.

Consul, P. C. and F. Famoye (1992). “Generalized Poisson Regression Model”. In: *Communication in Statistics – Theory and Methods* 21.1, pp. 89–109.

Consul, P. C. and G. C. Jain (1973). “A Generalization of the Poisson Distribution”. In: *Technometrics* 15.4, pp. 791–799.

Cox, D. R. (1962). *Renewal Theory.* Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Del Castillo, J. and M. Pérez-Casany (1998). “Weighted Poisson Distributions for Overdispersion and Underdispersion Situations”. In: *Annals of the Institute of Statistical Mathematics* 50.3, pp. 567–585.

Demétrio, C. G. B., J. Hinde, and R. A. Moral (2014). “Models for overdispersed data in entomology”. In: *Ecological modelling applied to entomology.* Springer, pp. 219–259.

Efron, B. (1986). “Double Exponential Families and Their Use in Generalized Linear Regression”. In: *Journal of the American Statistical Association* 84.395, pp. 709–721.

Gilbert, P. and R. Varadhan (2016). *numDeriv: Accurate Numerical Derivatives.* R package version 2016.8-1.

Hilbe, J. M. (2014). *Modeling Count Data.* New York: Cambridge University press.

Hinde, J. and C. G. B. Demétrio (1998). “Overdispersion: models and estimation”. In: *Computational Statistics & Data Analysis* 27.2, pp. 151–170.

Huang, A. (2017). “Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts”. In: *Statistical Modelling* 17.6, pp. 1–22.

Johnson, N. L., A. W. Kemp, and S. Kotz (2005). *Univariate Discrete Distributions.* 3rd edition. Series in Probability and Statistics. New Jersey: John Wiley & Sons.

Jørgensen, B. (1997). *The Theory of Dispersion Models.* Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Klakattawi, H. S., V. Vinciotti, and K. Yu (2018). “A Simple and Adaptive Dispersion Regression Model for Count Data”. In: *Entropy* 20.142.

Kokonendji, C. C. (2014). “Over- and Underdispersion Models”. In: ed. by N. Balakrishnan. John Wiley & Sons. Chap. 30, pp. 506–526.

Lindsey, J. K. (1996). *Parametric Statistical Inference.* New York: Oxford University Press.

Lord, D., S. D. Guikema, and S. R. Geedipally (2008). “Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes”. In: *Accident Analysis and Prevention* 40, pp. 1123–1134.

Luyts, M., G. Molenberghs, G. Verbeke, K. Matthijs, E. E. Ribeiro Jr, C. G. B. Demétrio, and J. Hinde (2018). “A Weibull-count approach for handling under- and overdispersed longitudinal/clustered data structures”. In: *Statistical Modelling* to appear.

Molenberghs, G., G. Verbeke, and C. G. B. Demétrio (2007). “An extended random-effects approach to modelling repetead, overdispersed count data”. In: *Lifetime Data Analysis* 13, pp. 513–531.

Nakagawa, T. and S. Osaki (1975). “The Discrete Weibull Distribution”. In: *IEEE Transactions on Reliability* 24.5, pp. 300–301.

Puig, P. and J. Valero (2006). “Count data distributions: some characterizations with applications”. In: *Journal of the American Statistical Association* 101.473, pp. 332–340.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

Ribeiro Jr, E. E., W. M. Zeviani, W. H. Bonat, C. G. B. Demétrio, and J. Hinde (2018). “Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data”. In: *arXiv (Statistics Applications and Statistics Methodology).*

Rigby, R. A. and D. M. Stasinopoulos (2005). “Generalized additive models for location, scale and shape (with discussion)”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.3, pp. 507–554.

Sellers, K. F. and D. S. Morris (2017). “Underdispersion models: Models that are “under the radar””. In: *Communication in Statistics – Theory and Methods* 46.24, pp. 12075–12086.

Sellers, K. F. and G. Shmueli (2010). “A flexible regression model for count data”. In: *Annals of Applied Statistics* 4.2, pp. 943–961.

Shmueli, G., T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright (2005). “A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.1, pp. 127–142.

Wedderburn, R. W. M. (1974). “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method”. In: *Biometrika* 61.3, p. 439.

Winkelmann, R. (1995). “Duration Dependence and Dispersion in Count-Data Models”. In: *Journal of Business & Economic Statistics* 13.4, pp. 467–474.

Winkelmann, R. (2008). *Econometric Analysis of Count Data.* 5th edition. Berlin, Heidelberg: Springer-Velag, p. 342.

Zamani, H. and N. Ismail (2012). “Functional Form for the Generalized Poisson Regression Model”. In: *Communication in Statistics – Theory and Methods* 41, pp. 3666–3675.

Zeviani, W. M., P. J. Ribeiro Jr, W. H. Bonat, S. E. Shimakura, and J. A. Muniz (2014). “The Gamma-count distribution in the analysis of experimental underdispersed data”. In: *Journal of Applied Statistics* 41.12, pp. 2616–2626.

Zou, Y., S. R. Geedipally, and D. Lord (2013). “Evaluating the double Poisson generalized linear model”. In: *Accident Analysis and Prevention* 59, pp. 497–505.

5 COM-POISSON MODELS WITH VARYING DISPERSION

ABSTRACT

There are many experiments that experimental conditions can have an effect on mean and dispersion. In this case, models with constant dispersion can produce a loss of efficiency. In this chapter, we propose an extension of the COM-Poisson model to jointly model the mean and the dispersion. This approach allows analyzing the data that exhibit under- and overdispersion depending on covariates. The estimation and inference are based on likelihood method. We carried out simulation studies to verify the finite sample properties of the maximum likelihood estimators. The results from our simulation study show that the maximum likelihood estimators are unbiased and consistent for both mean and dispersion regression parameters. The application of the COM-Poisson models with varying dispersion is illustrated with the analysis of two experimental datasets. The computational routines for fitting the models and the datasets are available in the appendix.

Keywords: COM-Poisson distribution, Double generalized linear models, GAMLSS, Varying dispersion.

5.1 Introduction

The standard Gaussian linear models are based on the assumption of variance homogeneity (Aitkin 1987). Generalized linear models relax this assumption by assuming the observations come from some distribution in the exponential family (Nelder and Wedderburn 1972). A key feature of exponential family distribution is the so-called mean-variance relationship, i.e. the variance is a deterministic function of the mean $\text{Var}(Y) = \phi V(\mu)$. The main examples are $V(\mu) = \mu(1 - \mu)$ for the binomial distribution, $V(\mu) = \mu$ for the Poisson distribution, $V(\mu) = \mu^2$ for the gamma distribution and $V(\mu) = \mu^3$ for the inverse-Gaussian distribution (McCullagh and Nelder 1989). This allows modeling specific heterogeneity by assuming an appropriate distribution. However, once the mean-variance relationship is specified, the variance is assumed to be known up to a constant of proportionality, the dispersion parameter~ϕ. To ensure more flexibility in the analysis of non-Gaussian heterogeneous data, we explore methods to model dispersion depending on covariates. We focus on count data only.

Modeling dispersion depending on covariates in the analysis of count data has not been much explored in the literature. The class of double generalized linear models (Smyth 1988; McCullagh and Nelder 1989; Smyth and Verbyla 1999) can be used to do it. This class was widely explored for continuous data. Smyth (1988) discussed this class using a double generalized linear model based on the gamma distribution. Adjusted likelihood methods for estimation and inference are presented by Smyth and Verbyla (1999). Paula (2013) discussed diagnostics for this class and used a assumed gamma distributed data as an application. Andersen and Bonat (2017) extended the double generalized linear model by considering compound Poisson distributions. Related to discrete data, Vieira et al. (2011) proposed a Bayesian analysis of the double generalized linear models for binomial data.

Another approach that has gained momentum in the last decade is the generalized additive models for location, shape, and scale (GAMLSS) (Rigby and Stasinopoulos 2005). This approach extends the generalized linear models in different directions: (i) the distribution for the response variable can be selected from a more general family; the only restriction is that the derivatives with respect to each parameter must be computable, (ii) all parameters of the distribution can be depended on covariates, and (iii) the systematic relationship between covariates and parameters can be parametric (linear predictor) or nonparametric (smooth) functions.

In this chapter, we propose to jointly model the mean and dispersion based on the COM-Poisson distribution. This approach is very similar to the GAMLSS, however, we develop and explore our own estimation methods. This approach allows data exhibit under- and overdispersion depending on experimental conditions.

This chapter is organized as follows. In Section 5.2, we briefly describe and discuss the double generalized linear models. The newly proposed COM-Poisson models is considered in Section 5.3. In Section 5.4, we present estimation and inference for the COM-Poisson regression model in a likelihood framework. Section 5.6 is devoted to illustrate the application of the COM-Poisson model with varying dispersion for the analysis of two data sets. We compare the results of COM-Poisson model with double generalized linear models and GAMLSS approach. Finally, concluding remarks close the chapter in Section 5.7. We provide an `R` implementation for fitting the COM-Poisson models with varying dispersion, together with the analyzed data sets, ins the package `cmpreg`. Illustrative codes are presented in the appendix.

5.2 Double generalized linear models

Following Smyth (1988) and McCullagh and Nelder (1989, chap. 10), the so-called double generalized linear models (DGLM) can be used to model mean and dispersion jointly by considering two linked generalized linear models (GLM). Let y_i, $i = 1, 2, \dots, n$ be a set of independent observations of Y_i and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{np})^\top$ a vector of known covariates. The standard GLMs assume that

$$\mathrm{E}(Y_i) = \mu_i, \quad g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad \mathrm{Var}(Y_i) = \phi V(\mu_i),$$

where $g(.)$ is a suitable link function and $V(.)$ is the variance function. The DGLMs generalize the GLMs to allow the dispersion parameter varying across observations, $\mathrm{Var}(Y_i) = \phi_i V(\mu_i)$. The model for dispersion is specified by

$$\mathrm{E}(d_i) = \phi_i, \quad h(\phi_i) = \xi_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma} \quad \text{and} \quad \mathrm{Var}(d_i) = \tau V_d(\phi_i),$$

where d_i is a suitable statistic chosen as a measure of dispersion, $h(.)$ is the dispersion link function and $V_d(.)$ is the dispersion variance function. The commom choice for d_i is the deviance components from the mean model. Based on asymptotic distribution of d_i, a gamma model with $V_d(\phi_i) = 2\phi_i^2$ is a suitable natural choice for the dispersion model.

The parameter estimation proposed is based on a two-step iterative algorithm: (i) holding $\boldsymbol{\gamma}$ fixed we obtain the dispersion components and estimate $\boldsymbol{\beta}$; and (ii) fixing the estimated value of $\boldsymbol{\beta}$, we obtain the deviance components and estimate $\boldsymbol{\gamma}$. These two steps are then alternated until convergence.

The idea of using a interlinked pair of generalized linear models for allowing mean and dispersion effects depend on covariates was first put forward by Pregibon (1984) and further developed by different authors with different names. Smyth (1988) called it generalized linear models with varying dispersion, while McCullagh and Nelder (1989), at the same time, named this approach by joint modelling of mean and dispersion. Nowadays, the most used name to refer to this approach is double generalized linear models, introduced by Smyth and Verbyla (1999).

Another approach that has become popular for modeling mean and dispersion is the class of generalized additive models for location, scale, and shape (GAMLSS) (Rigby and Stasinopoulos 2005). It is important to note that the DGLMs for discrete data are not a particular case of GAMLSS. Although both can model mean and dispersion jointly in a similar way, the GAMLSS require a two-parameter distribution for the response. In this chapter, we compare the proposed COM-Poisson model varying dispersion with a GAMLSS based on the double Poisson distribution (Efron 1986; Zou, Geedipally, and Lord 2013) and a DGLM Poisson.

5.3 Modeling mean and dispersion COM-Poisson

The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution that can deal with under-, over- and equidispersion (Shmueli et al. 2005; Sellers and Shmueli 2010). The probability mass function of the COM-Poisson distribution is given by

$$\Pr(Y = y) = \frac{\lambda^y}{(y!)^\nu \mathrm{Z}(\lambda, \nu)}, \quad \text{where} \quad \mathrm{Z}(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}. \tag{5.1}$$

The series $\mathrm{Z}(\lambda, \nu)$ is a normalizing constant that cannot be expressed in closed form unless for special cases. The parameters λ and ν can be seen in terms of the ratios of successive probabilities, $\Pr(Y = y - 1)/\Pr(Y = y) = y^\nu/\lambda$.

The parameter ν is the dispersion parameter with a clear interpretation; overdispersion for $0 < \nu < 1$ and underdispersion for $\nu > 1$. When $\nu = 1$, the Poisson distribution results as a special case. On the other hand, the parameter λ has no clear interpretation, unless for $\nu = 1$, and it is strongly related to ν. To circumvent this problem, Ribeiro Jr et al. (2018) propose a reparameterization of the COM-Poisson distribution to provide an approximate mean parameter. By introducing the new parameter $\mu > 0$,

$$\mu = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad \Rightarrow \quad \lambda = \left(\mu + \frac{(\nu - 1)}{2\nu}\right)^\nu,$$

the authors showed that the new parametrization has good properties for estimation and inference and proposed a regression model on approximated mean. In this paper, we propose to allow both μ and ν parameters depending on covariates.

Let y_i, $i = 1, 2, \ldots, n$ be a set of independents realizations of Y_i following a COM-Poisson distribution with parameters μ_i and ν_i. The proposed COM-Poisson varying dispersion model assumes

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad \xi_i = h(\nu_i) = \boldsymbol{z}_i^\top \boldsymbol{\gamma},$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)^\top$ are the parameters to be estimated, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{np})^\top$ and $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{nq})^\top$ are vectors of known covariates, and $g(.)$ and $h(.)$ are suitable link functions. We use the logarithmic link function for both mean and dispersion.

The COM-Poisson model with varying dispersion is quite similar to the double generalized linear approach, both extend the generalized linear models to model the dispersion as well as the mean. However, COM-Poisson with varying dispersion is a fully parametric model that has the advantages of allowing predictions for probabilities and generalizations such as random effects and modeling of censored data.

5.4 Estimation and inference

To fit COM-Poisson models with varying dispersion, we use the maximum likelihood estimation method. The log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ parameters is given by

$$\ell = \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \nu_i \log\left(\mu_i + \frac{\nu_i - 1}{2\nu_i}\right) - \nu_i \log(y_i) - \log[Z(\mu_i, \nu_i)] \right\}, \tag{5.2}$$

where $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, $\nu_i = \exp(\boldsymbol{z}_i^\top \boldsymbol{\gamma})$, and $Z(\mu_i, \nu_i)$ is a normalizing constant computed for the parameters μ_i and ν_i.

Parameter estimation requires the numerical maximization of Equation (5.2). Since the derivatives of ℓ cannot be obtained in closed forms, we compute them by central finite differences using the Richardson method as implemented in package `numDeriv` (Gilbert and Varadhan 2016) for the statistical software `R` (R Core Team 2018). Based on the orthogonality property between $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_i$ (see Ribeiro Jr et al. 2018), we set two strategies to obtain the maximum likelihood estimates:

(a) the *joint* strategy, where we obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ jointly by maximizing $\ell(\boldsymbol{\theta}; \boldsymbol{y})$; and

(b) the *two-stages* strategy, where we fix $\hat{\boldsymbol{\beta}}$ at the maximum likelihood estimates of Poisson model and obtain $\hat{\boldsymbol{\gamma}}$ by maximizing $\ell(\boldsymbol{\gamma}; \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \boldsymbol{y})$ and then estimate the Hessian matrix at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$.

Although we have an additional step for estimating the Hessian matrix, the strategy (b) tends to be faster than (a), because of the maximization is performed in $\mathbb{R}^q$ whereas, for strategy (a), the estimation is an optimization problem in $\mathbb{R}^{p+q}$ space.

Standard errors for the parameter estimates are obtained based on the observed information matrix. Let the derivatives be computed at the maximum likelihood estimates. The variance and covariance matrix of the maximum likelihood estimators may be expressed as

$$\boldsymbol{V}_\theta = \begin{pmatrix} -\partial\ell^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top & -\partial\ell^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^\top \\ -\partial\ell^2/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^\top & -\partial\ell^2/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{V}_\beta & \boldsymbol{V}_{\beta\gamma} \\ \boldsymbol{V}_{\gamma\beta} & \boldsymbol{V}_\gamma \end{pmatrix}.$$

Based on the asymptotic distribution of the maximum likelihood estimators, we obtain the variances for $\hat{\eta}_i$ and $\hat{\xi}_i$ using delta method, $\text{Var}(\hat{\eta}_i) = \boldsymbol{x}_i^\top \boldsymbol{V}_{\beta|\gamma} \boldsymbol{x}_i$ and $\text{Var}(\hat{\xi}_i) = \boldsymbol{z}_i^\top \boldsymbol{V}_{\gamma|\beta} \boldsymbol{z}_i$, where $\boldsymbol{V}_{\beta|\gamma} = \boldsymbol{V}_\beta - \boldsymbol{V}_{\beta\gamma}\boldsymbol{V}_\gamma^{-1}\boldsymbol{V}_{\gamma\beta}$ and $\boldsymbol{V}_{\gamma|\beta} = \boldsymbol{V}_\gamma - \boldsymbol{V}_{\gamma\beta}\boldsymbol{V}_\beta^{-1}\boldsymbol{V}_{\beta\gamma}$. Since $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ were found to be nearly orthogonal, $\boldsymbol{V}_{\beta\gamma} = \boldsymbol{V}_{\gamma\beta}^\top \approx \boldsymbol{0}$, hence $\boldsymbol{V}_{\beta|\gamma} \approx \boldsymbol{V}_\beta$ and $\boldsymbol{V}_{\gamma|\beta} \approx \boldsymbol{V}_\gamma$, in other words, inference based on the conditional log-likelihood and inference based on the marginal log-likelihood is the same.

Confidence intervals for μ_i and ν_i are obtained by transforming the confidence intervals for η_i and ξ_i. We implemented the two strategies for fitting the COM-Poisson models and the methods for computing the confidence intervals in the `cmpreg`[1] package for the software `R`.

5.5 Simulation study

In this section, we present a simulation study designed to assess the properties of the maximum likelihood estimators using the proposed (a) joint and (b) two-stages strategies. We compared the estimation strategies by maximized log-likelihoods and fitting times.

We considered average counts varying from 3 to 30 and average dispersion varying from 0.3 to 4 arising from regression models with a continuous (x_1) and a categorical (x_2) covariate. The continuous covariate was generated as a linearly increasing sequence from 0 to 1 with length equal to the sample size. Similarly, the categorical covariate was generated as a sequence of three values each one repeated n/3 times (rounding up when required), where n denotes the sample size. The parameter μ and ν of the reparametrized COM-Poisson random variable is given by $\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \beta_{22} x_{22}$ and $\log(\nu) = \gamma_0 + \gamma_1 x_1 + \gamma_{21} x_{21} + \gamma_{22} x_{22}$, respectively, where x_{21} and x_{22} are dummy representing the levels of x_2. The regression coefficients were fixed at the values, $\beta_0 = 2, \beta_1 = 0.5, \beta_{21} = 0.8$ and $\beta_{22} = -0.8$ for the mean, and $\gamma_0 = -0.7, \gamma_1 = 1.3, \gamma_{21} = 0.8$ and $\gamma_{22} = -0.6$ for the dispersion. Figure 5.1 shows the parameter space evaluated by this simulation study. With this simulation design, we are able to assess the properties of the parameter estimators in situation of high and low counts and equi-, under-, and overdispersion.

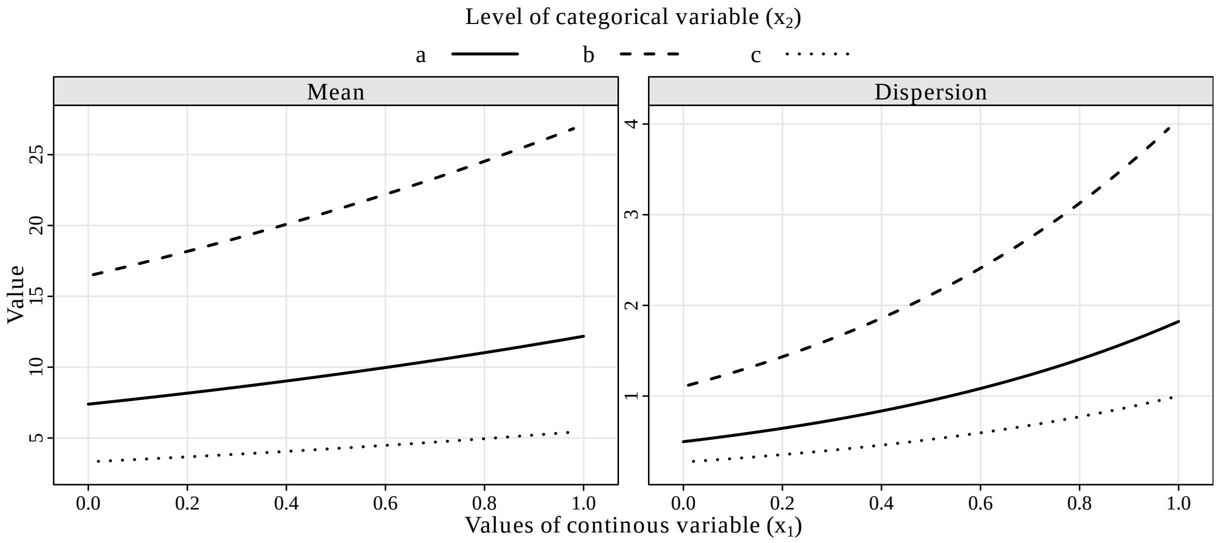

Figure 5.1. Values for the mean (left) and for the dispersion (right) according to the regression models adopted.

In order to check the consistency of the estimators we considered four different sample sizes: $50, 100, 300$ and 1000; generating 1000 data sets in each case. In Figure 5.2, we show the bias of the estimators along with the confidence intervals calculated as average bias plus and minus 1.96 times the average standard error. The scales are standardized for each parameter by

[1] Available on GitHub https://github.com/jreduardo/cmpreg

dividing the average bias by the average standard error obtained for the sample of size 50. The results show that for both estimation methods the expected bias and the standard error tend to 0 as the sample size increases. Thus, this shows the unbiasedness, consistency, and symmetry of the empirical distribution of the maximum likelihood estimators using both strategies.

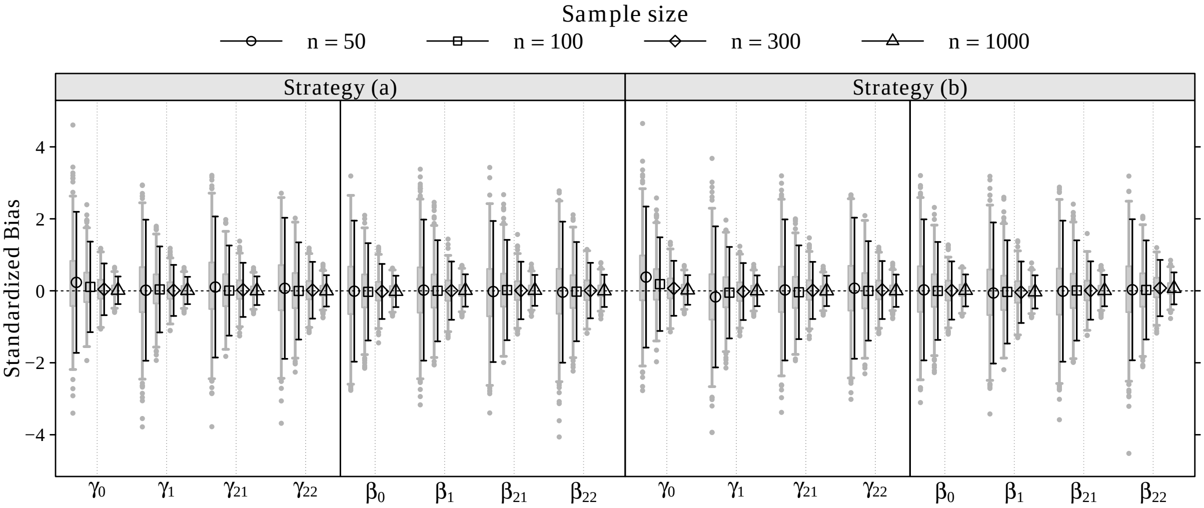

Figure 5.2. Distributions of standardized bias (gray box-plots) and average with confidence intervals (black segments) by different sample sizes.

In order to compare the proposed strategies, we compute the log-likelihood function at the maximum likelihood estimates for each simulated data set. The results are shown in Figure 5.3. The maximized log-likelihood obtained using the strategy (b) is always smaller than that obtained considering the strategy (a). This is expected since in strategy (b) we do not refit the mean regression parameters. However, the likelihoods values are similar between the two strategies, showing that the estimates of the Poisson model are very close to the genuine mean regression maximum likelihood estimates of the COM-Poisson.

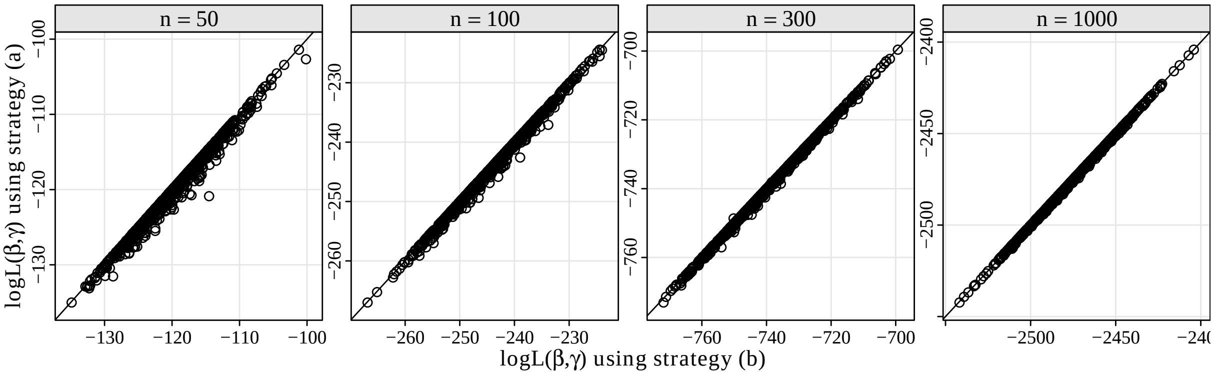

Figure 5.3. Scatter plots of maximized log-likelihood obtained using the strategy (a) *vs.* those obtained using the strategy (b).

The computational times to fit the models using the both strategies are shown in Table 5.1. As highlighted in the previous section, we expect the strategy (b) to be faster, since the maximization is done in a smaller dimension. The results show that the estimation by using the strategy (a) takes, on average, around 30% more time than using the strategy (b).

Table 5.1. Average and quartiles of the time to fit the COM-Poisson models using strategies (a) and (b), in seconds, and ratios between the averages.

Sample size	Strategy (a)			Strategy (b)			
	Average	1st quartile	3rd quartile	Average	1st quartile	3rd quartile	Ratio
50	1.01	0.98	1.07	0.77	0.77	0.82	1.31
100	1.80	1.64	1.94	1.37	1.26	1.50	1.31
300	3.89	3.76	4.78	2.98	2.92	3.60	1.30
1000	14.24	12.37	15.87	11.13	9.70	12.47	1.28

5.6 Applications and discussion

In this section, we shall present two data analyzes to illustrate the application of the COM-Poisson model with varying dispersion. We compare the results of the COM-Poisson model with the double generalized linear models approach, fitted using `dglm::dglm(..., family = poisson)`, and double Poisson model, fitted using `gamlss::gamlss(..., family = DPO)`. The data sets and `R` codes used in this section are available in the appendix.

5.6.1 Analysis of nitrofen experiment

This dataset comes from a completely randomized experiment where the number of live offspring of a species of zooplankton were recorded for different doses of nitrofen. These data are presented in Section 2.3 and were analyzed by Ribeiro Jr et al. (2018) using the COM-Poisson model with constant dispersion.

Here, to analyze the number of live offspring under different doses of nitrofen, we consider the following linear predictors for the dispersion

$$\begin{array}{ll}\text{Constant:} & \log(\nu_{ij}) = \gamma_0,\\ \text{Linear:} & \log(\nu_{ij}) = \gamma_0 + \gamma_1 \mathtt{x}_{1i},\\ \text{Quadratic:} & \log(\nu_{ij}) = \gamma_0 + \gamma_1 \mathtt{x}_{1i} + \gamma_2 \mathtt{x}_{2i}^2 \text{ e}\\ \text{Cubic:} & \log(\nu_{ij}) = \gamma_0 + \gamma_1 \mathtt{x}_{1i} + \gamma_2 \mathtt{x}_{2i}^2 + \gamma_3 \mathtt{x}_{3i}^3,\end{array}$$

where i and j refers to the nitrofen concentration level (`dose`) and to the replicates, respectively. The covariates $\mathtt{x}_1$, $\mathtt{x}_2$ and $\mathtt{x}_3$ are the orthogonal linear, quadratic and cubic polynomials of the `dose`. To the mean, we consider a raw cubic polynomial of `dose`.

In Table 5.2, some goodness-of-fit measures (minus two times the maximized log-likelihood and Akaike information criterion) are presented together with corresponding likelihood ratio tests between nested models. The results show a clear evidence of that the dispersion of the counts is at least linearly dependent on the nitrofen concentration level. Consider the COM-Poisson model, the difference of deviances of the models considering the quadratic and linear predictors is greater than the difference obtained for double Poisson and double GLM Poisson models.

Table 5.2. Nitrofen data: goodness-of-fit measures (deviance and AIC) and model comparisons.

		COM-Poisson			
	G.l	Deviance	AIC	χ^2	$\Pr(>\chi^2)$
Constant	45	288.127	298.127		
Linear	44	274.111	286.111	14.0164	0.0002
Quadratic	43	270.493	284.493	3.6179	0.0572
Cubic	42	269.503	285.503	0.9898	0.3198

		double Poisson (`GAMLSS`)			
	G.l	Deviance	AIC	χ^2	$\Pr(>\chi^2)$
Constant	45	288.181	298.181		
Linear	44	273.530	285.530	14.6512	0.0001
Quadratic	43	271.204	285.204	2.3256	0.1273
Cubic	42	269.201	285.201	2.0031	0.1570

		double GLM Poisson (`DGLM`)			
	G.l	Deviance	AIC	χ^2	$\Pr(>\chi^2)$
Constant	45	287.690	297.690		
Linear	44	272.634	284.634	15.0565	0.0001
Quadratic	43	269.997	283.997	2.6368	0.1044
Cubic	42	268.165	284.165	1.8317	0.1759

Deviance is computed as minus twice log-likelihood

Table 5.3. Nitrofen data: Parameter estimates (Est) and standard errors (SEs) for the fitted double regression COM-Poisson model.

	Estimates (Standard Errors)			
Parameter	Constant	Linear	Quadratic	Cubic
Mean				
β_0	3.4769 (0.0541)[a]	3.4778 (0.0283)[a]	3.4777 (0.0339)[a]	3.4776 (0.0360)[a]
β_1	−0.0879 (0.1943)	−0.1055 (0.1424)	−0.1167 (0.1322)	−0.1153 (0.1283)
β_2	0.1547 (0.1731)	0.1747 (0.1492)	0.1917 (0.1375)	0.1886 (0.1318)
β_3	−0.0976 (0.0396)[a]	−0.1028 (0.0380)[a]	−0.1082 (0.0368)[a]	−0.1069 (0.0350)[a]
Dispersion				
γ_0	0.0474 (0.2047)	0.2948 (0.2112)	0.2437 (0.2589)	0.3532 (0.2268)
γ_1	–	−5.2441 (1.3630)[a]	−7.0024 (2.3054)[a]	−5.7301 (1.8440)[a]
γ_2	–	–	−3.9807 (2.4439)	−2.9175 (1.9045)
γ_3	–	–	–	1.5221 (1.4119)

Est (SE)[a] indicates $|\text{Est/SE}| > 1,96$.

The estimated parameters, their associated standard errors and individual Wald tests are presented in Table 5.3 for COM-Poisson model. We fitted the models using the strategy (a), for that reason the parameter estimates for mean structure are slightly different for the different predictors. For the dispersion structure, there is no evidence to keep the quadratic term, the Wald's statistics is equals to -1.63 (p-value $= 0.1034$).

Figure 5.4 presents the fitted mean values $\hat{\mu}_i$ and fitted dispersion values $\hat{\nu}_i$ across doses between 0 and $3.1\mu\text{g}/10^2$litre for the four different linear predictors adopted to dispersion.

When we consider the dispersion constant, the model indicates equidispersion ($\nu = 1$). However, when we relax this assumption, it is clear that the dispersion change across nitrofen levels. In particular, all models show that around 2μg/10^2litre the number of live offspring change from under- to overdispersed.

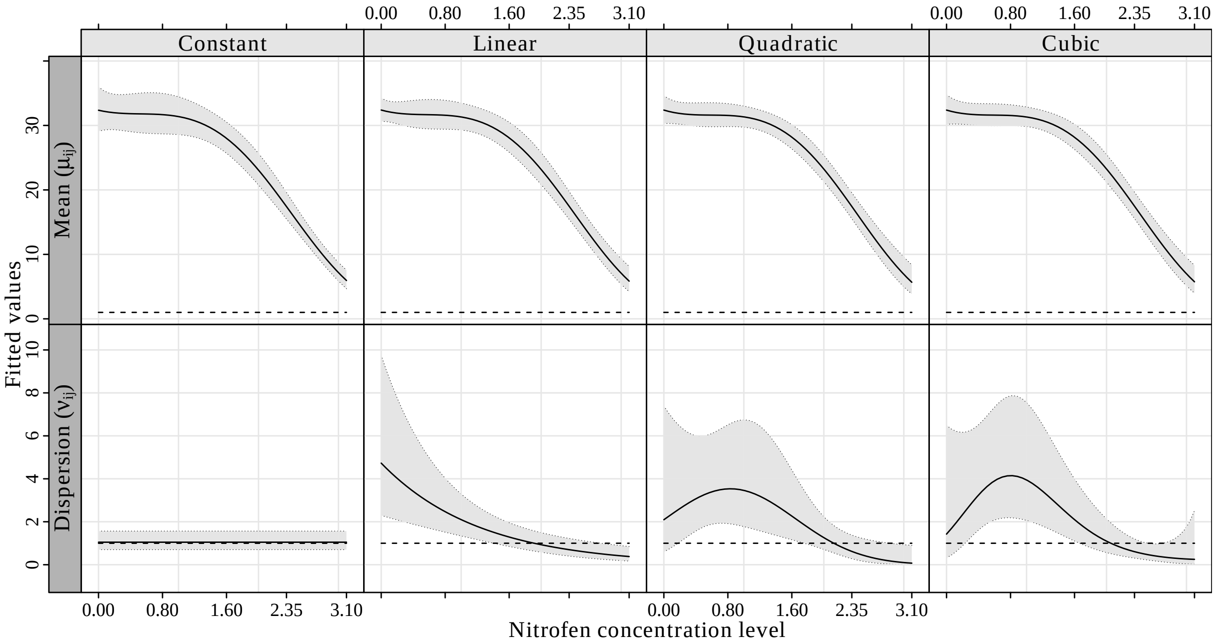

Figure 5.4. Fitted mean and dispersion values with 95% confidence intervals for the considered linear predictors.

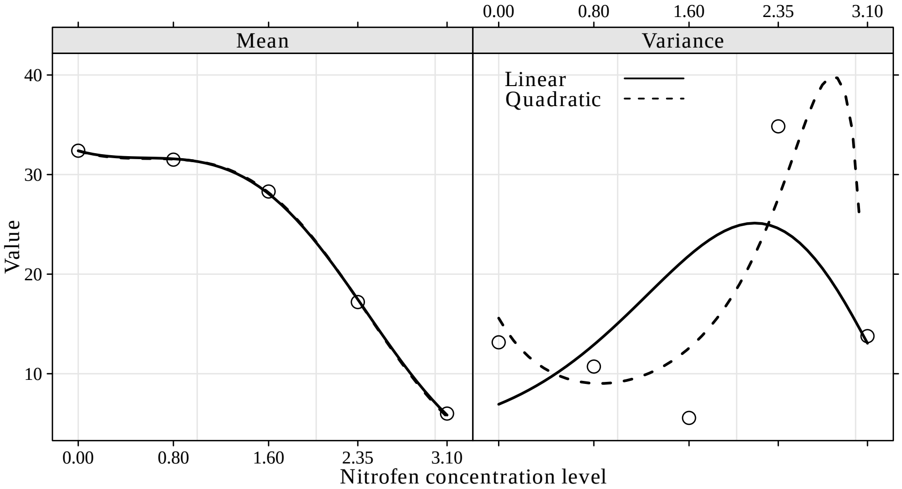

Figure 5.5. Expected values and variances obtained from the fitted models with linear and quadratic predictors for the dispersion. The points are the sample means and sample variances, respectively.

The expected values and variances for nitrofen doses between 0 and 3.1μg/10^2litre, obtained from the fitted models with linear and quadratic predictors for the dispersion, are presented in Figure 5.5. Note that although the polynomial adopted for the mean is not the saturated one, the fitted curve almost interpolates the sample averages. The predicted variances show interesting behavior. Note that these curves are nonlinear functions of μ_i and ν_i, so a linear predictor of ν_i does not imply a linear curve here. The model with a linear predictor for the dispersion does not fit the sample variances for the doses 1.6 and 2.35μg/10^2litre well. The model with a quadratic predictor fits these variances better, however, it produces rather strange behavior for values greater than 2.35μg/10^2litre.

5.6.2 Analysis of soybean experiment

In this second application, we analyze the data resulting from a 5×3 factorial experiment conducted in a randomized complete block design, see Section 2.2. In this experiment, we have two counting responses recorded, number of grains and number of pods. Here we focus on analyzing the number of pods and consider the following predictors for mean and dispersion

$$\begin{aligned}\log(\mu_{ijk}) &= \beta_0 + \kappa_i + \tau_j + (\beta_1 + \delta_j)\texttt{K}_k + \beta_2^2\texttt{K}_k^2\\ \log(\nu_{ijk}) &= \alpha_j + \gamma_1\texttt{K}_k,\end{aligned}$$

where i, j, and k varies according to block, moisture level, and potassium fertilization. Note that for the dispersion, the α_j is the logarithm of dispersion parameter when the potassium dose is 0 (without an intercept term).

In Table 5.4, we present the goodness-of-fit measures (minus twice log-likelihood and Akaike information criterion) together with the corresponding likelihood ratio tests between the current model (moisture + K) and the reduced models for dispersion: $\gamma_1 = 0$ (moisture effect) and $\alpha_j = \alpha$ (constant). The log-likelihood achieved for the COM-Poisson, double Poisson and double generalized linear Poisson models are practically the same for all three predictors for dispersion. The results show a significant improvement in varying the dispersion for each moisture level and do not present a clear relation between the dispersion and the potassium doses when moisture is already included in the linear predictor.

Table 5.5 gives the parameter estimates and associated standard errors for the three different predictors. It seems that the mean structure is well specified since the Wald tests lead to the same conclusions for the different dispersion structures. The likelihood ratio tests indicate that we have different dispersion levels for each moisture level. In particular, considering the reduced model $\gamma_1 = 0$, the fitted dispersions values for each moisture level are $\hat{\nu}_{37.5\%} = 0.701$, $\hat{\nu}_{50.0\%} = 0.742$, and $\hat{\nu}_{63.5\%} = 1.99$. The point estimates indicate overdispersion for the two first and underdispersion for the latter.

The 95% confidence intervals for the dispersion components are shown in Figure 5.6. The intervals are obtained by working with the profile deviance function with respect to each α_j. The results indicate neither under- nor overdispersion for moisture levels 37.5% and 50%. For the level 63.5%, there is evidence for underdispersion, although the lower bound is close to zero.

The observed and fitted counts for each moisture level with confidence intervals are shown in Figure 5.7, along with the optimum doses. The results show that potassium doses at 104.21, 132.63 and 132.63mg dm^3 lead to the expected maximum number of pods for the moisture level 37.5%, 50%, and 62.5%, respectively. In terms of the number of pods, there is evidence that fertilization with potassium can compensate the water deficit in soybean culture.

Table 5.4. Soybean data (pods): goodness-of-fit measures (deviance and AIC) and model comparisons.

		COM-Poisson			
	Df	Deviance	AIC	χ^2	Pr(> χ^2)
Constant	62	534.232	558.232		
Moisture	60	527.866	555.866	6.3658	0.0415
Moisture + K	59	525.848	555.848	2.0181	0.1554
		double Poisson (`GAMLSS`)			
	Df	Deviance	AIC	χ^2	Pr(> χ^2)
Constant	62	534.231	558.231		
Moisture	60	527.875	555.875	6.3566	0.0417
Moisture + K	59	525.874	555.874	2.0005	0.1572
		double GLM Poisson (`DGLM`)			
	Df	Deviance	AIC	χ^2	Pr(> χ^2)
Constant	62	534.042	558.042		
Moisture	60	527.677	555.677	6.3654	0.0415
Moisture + K	59	525.663	555.663	2.0137	0.1559

Deviance is computed as $-2\times$logLik.

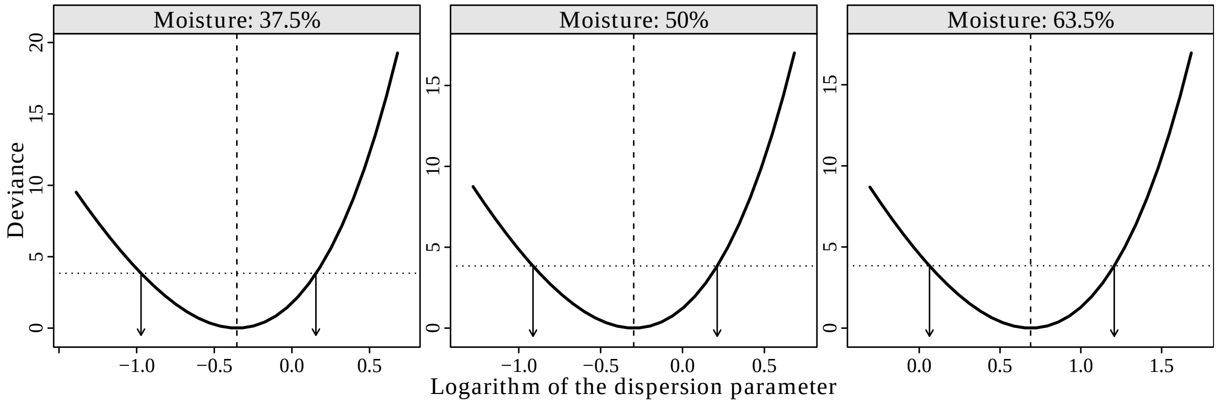

Figure 5.6. Profile deviation for each dispersion component and their respective 95% confidence intervals.

In this analysis, we considered a quadratic predictor for the number of pods along the potassium doses under which we could find the optimal doses for each moisture level. However, in practice, this is not a reasonable behavior as the number of pods decreases with increasing doses of potassium. Here, a nonlinear predictor with a plateau or asymptote can be very useful with direct interpretations, but is was not explored here.

5.7 Final remarks

In this paper, COM-Poisson models with varying dispersion were proposed for jointly modeling mean and dispersion in the analysis of counts with different levels of dispersion. This

Table 5.5. Soybean data (pods): Parameter estimates (Est) and standard errors (SEs) for the fitted double regression COM-Poisson model.

	Estimates (Standard Errors)		
Parameter	Constant	Moisture	Moisture + K
Mean			
β_0	4.0062 (0.0544)^{a}	4.0112 (0.0576)^{a}	4.0273 (0.0632)^{a}
κ_{II}	−0.0293 (0.0426)	−0.0320 (0.0381)	−0.0398 (0.0359)
κ_{III}	−0.0727 (0.0431)	−0.0991 (0.0393)^{a}	−0.1097 (0.0379)^{a}
κ_{IV}	−0.1254 (0.0437)^{a}	−0.1452 (0.0403)^{a}	−0.1573 (0.0380)^{a}
κ_{V}	−0.1037 (0.0446)^{a}	−0.0960 (0.0411)^{a}	−0.1016 (0.0394)^{a}
$\tau_{50.0\%}$	0.1573 (0.0583)^{a}	0.1571 (0.0659)^{a}	0.1557 (0.0723)^{a}
$\tau_{63.5\%}$	0.1730 (0.0582)^{a}	0.1731 (0.0562)^{a}	0.1758 (0.0610)^{a}
β_1	0.5851 (0.0902)^{a}	0.5976 (0.0858)^{a}	0.5596 (0.0898)^{a}
$\delta_{50.0\%}$	0.1469 (0.0552)^{a}	0.1471 (0.0625)^{a}	0.1484 (0.0600)^{a}
$\delta_{63.5\%}$	0.1398 (0.0553)^{a}	0.1407 (0.0539)^{a}	0.1374 (0.0511)^{a}
β_2	−0.2683 (0.0434)^{a}	−0.2751 (0.0387)^{a}	−0.2545 (0.0397)^{a}
Dispersion			
$\alpha_{37.5\%}$	−0.0826 (0.1652)	−0.3546 (0.3145)	−0.7023 (0.4067)
$\alpha_{50.0\%}$	–	−0.2979 (0.2982)	−0.6413 (0.3979)
$\alpha_{63.5\%}$	–	0.6893 (0.3023)^{a}	0.4615 (0.3459)
γ_1	–	–	0.4314 (0.2983)

Est (SE)^{a} indicates $|\text{Est/SE}| > 1,96$.

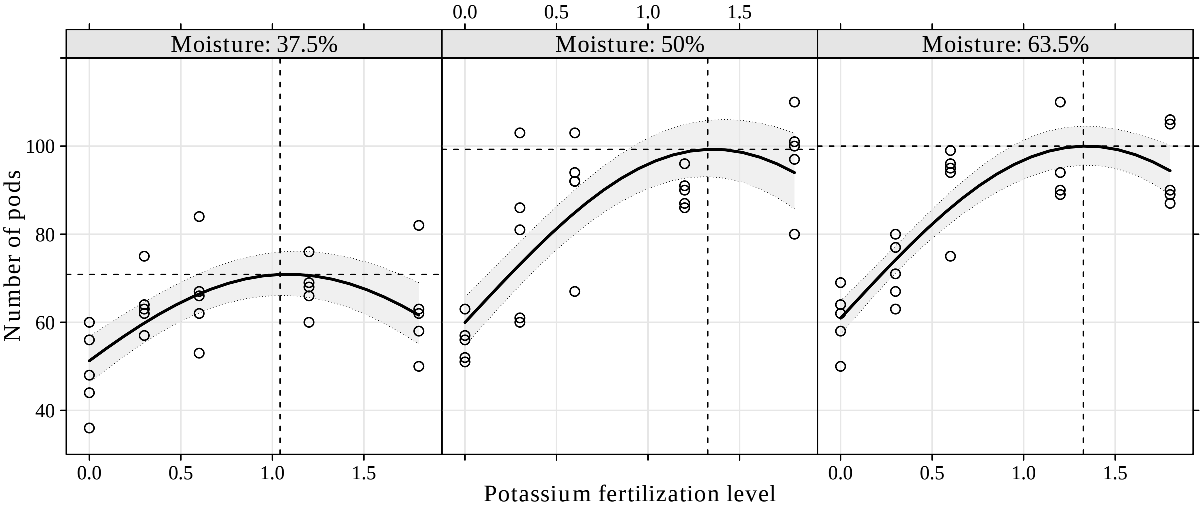

Figure 5.7. Fitted curves for the number of pods with 95% confidence intervals. The dotted lines indicate the optimal doses and the corresponding expected number of pods.

class of models allows modeling the dispersion depending on covariates. The parameters are estimated by the maximum likelihood method and inferences are made based on the asymptotic distributions of the estimators.

We carried out a simulation study to assess the properties of the maximum likelihood estimators obtained by two proposed strategies. The results of our simulation study suggested that the maximum likelihood estimators for the mean and dispersion regression parameters are unbiased and consistent. The comparison of the strategies for maximum likelihood estimation indicate that using the strategy (a) (maximization of conditional likelihood) reduces the fitting times and leads to inferences quite similar to the strategy (b) (maximization of full likelihood). However, studies regarding the coverage rate of confidence intervals and the evaluation of type 1 errors in hypothesis testing are necessary.

The methodology is applied to analyze two count datasets obtained from planned experiments. We compared the COM-Poisson models with double generalized linear models and GAMLSS approach. The proposal presented improvements in terms of fitting data when compared to the conventional COM-Poisson with constant dispersion and was competitive with the double generalized linear models and GALMSS based on double Poisson distribution.

For future work, simulation studies are necessary in order to evaluate the robustness of the model. In addition, as an improvement in the case study analysis, considering nonlinear predictors for the mean can be useful in order to avoid the use of high-order polynomials and to impose biological restrictions.

References

Aitkin, M. (1987). “Modelling Variance Heterogeneity in Normal Regression Using GLIM”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36.3, pp. 332–339.

Andersen, D. A. and W. H. Bonat (2017). “Double Generalized Linear Compound Poisson models to Insurance Claims Data”. In: *Electronic Journal of Applied Statistical Analysis* 10.2.

Efron, B. (1986). “Double Exponential Families and Their Use in Generalized Linear Regression”. In: *Journal of the American Statistical Association* 84.395, pp. 709–721.

Gilbert, P. and R. Varadhan (2016). *numDeriv: Accurate Numerical Derivatives.* R package version 2016.8-1.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models.* 2nd edition. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Nelder, J. A. and R. W. M. Wedderburn (1972). “Generalized Linear Models”. In: *Journal of the Royal Statistical Society. Series A (General)* 135, pp. 370–384.

Paula, G. A. (2013). “On diagnostics in double generalized linear models”. In: *Computational Statistics & Data Analysis* 68, pp. 44–51.

Pregibon, D. (1984). “Review: P. McCullagh, J. A. Nelder, Generalized Linear Models””. In: *The Annals of Statistics* 12.4, pp. 1589–1596.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

Ribeiro Jr, E. E., W. M. Zeviani, W. H. Bonat, C. G. B. Demétrio, and J. Hinde (2018). “Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data”. In: *arXiv (Statistics Applications and Statistics Methodology).*

Rigby, R. A. and D. M. Stasinopoulos (2005). “Generalized additive models for location, scale and shape (with discussion)”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.3, pp. 507–554.

Sellers, K. F. and G. Shmueli (2010). “A flexible regression model for count data”. In: *Annals of Applied Statistics* 4.2, pp. 943–961.

Shmueli, G., T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright (2005). “A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.1, pp. 127–142.

Smyth, G. K. (1988). “Generalized Linear Models with Varying Dispersion”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 51.1, pp. 47–60.

Smyth, G. K. and A. P. Verbyla (1999). “Adjusted likelihood methods for modelling dispersion in generalized linear models”. In: *Environmetrics* 10.6, pp. 695–709.

Vieira, A. M., R. A. Leandro, C. G. Demétrio, and G. Molenberghs (2011). “Double generalized linear model for tissue culture proportion data: a Bayesian perspective”. In: *Journal of Applied Statistics* 38.8, pp. 1717–1731.

Zou, Y., S. R. Geedipally, and D. Lord (2013). “Evaluating the double Poisson generalized linear model”. In: *Accident Analysis and Prevention* 59, pp. 497–505.

6 FINAL CONSIDERATIONS

In this thesis, we aimed to explore and extend statistical models for the analysis of dispersed count data. This was motivated by five datasets from planned experiments in the agricultural and biological contexts. Initially, we planned to work with the COM-Poisson distribution and propose a new parameterization only, reported in Chapter 3. However, the study of the alternative models for the analysis of count data led us to deepen the comparison of these models in order to find in which situations each one would fit better, giving rise to Chapter 4. Finally, motivated by the nitrofen data, we noticed that it was necessary to relax the assumption of constant dispersion and we developed the class of COM-Poisson with varying dispersion (Chapter 5).

The results of this thesis will be published in statistical journals to report our findings to the scientific community. The proposed new parametrization of the COM-Poisson model in Chapter 3 was submitted to Statistical Modelling Journal in February 2018 (Ribeiro Jr et al. 2018). We also contributed to the paper by Luyts et al. (2018), where some findings and extensions were reported in Chapter 4.

All computational routines and datasets used in this thesis are available in the respective GitHub repositories[1]. We used mostly the statistical software `R` (R Core Team 2018) and we are proud to have put some effort into organizing the functions into two `R` packages: `flexcm` and `cmpreg`. These packages contribute to the reproducible research as well as facilitate readers who wish to make further extensions and/or comparisons.

There are many possibilities to further work. As highlighted in the data analysis, there are many datasets that present a nonlinear response over some covariate (time, dose, etc.). Therefore, a natural extension is to allow nonlinear predictors for dispersed count data. A initial reference may be the generalized nonlinear models by Turner and Firth (2007). Besides that, it is very common the collection of correlated data, such as resulting from multivariate, longitudinal, spatial, and clustered designs (Molenberghs and Verbeke 2005). Such designs motivate generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993), where random effects are included at the group levels introducing correlation between observations. So, to model the correlation in dispersed count we can incorporate random effects in the flexible models presented in Chapter 3. A related work can be found in Brooks et al. (2017), Lee and Nelder (2006) and Rigby and Stasinopoulos (2005). Finally, taking advantage of the fully parametric specification, censoring can be incorporated easily in these models.

References

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Mächler, and B. M. Bolker (2017). “glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling”. In: *The R Journal* 9.2, pp. 378–400.

[1] See all repositories in https://github.com/jreduardo

Lee, Y. and J. A. Nelder (2006). “Double hierarchical generalized linear models (with discussion)”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 55, pp. 139–185.

Luyts, M., G. Molenberghs, G. Verbeke, K. Matthijs, E. E. Ribeiro Jr, C. G. B. Demétrio, and J. Hinde (2018). “A Weibull-count approach for handling under- and overdispersed longitudinal/clustered data structures”. In: *Statistical Modelling* to appear.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data.* Series in Statistics. New York: Springer.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

Ribeiro Jr, E. E., W. M. Zeviani, W. H. Bonat, C. G. B. Demétrio, and J. Hinde (2018). “Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data”. In: *arXiv (Statistics Applications and Statistics Methodology).*

Rigby, R. A. and D. M. Stasinopoulos (2005). “Generalized additive models for location, scale and shape (with discussion)”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.3, pp. 507–554.

Turner, H. and D. Firth (2007). *Generalized nonlinear models in R: An overview of the gnm package.* Tech. rep. ESRC National Centre for Research Methods.

APPENDIX

Appendix A: R packages and computational routines

In this dissertation, we implemented two packages, `cmpreg`[2] and `flexcm`[3] for the statistical software R (R Core Team 2018). The packages are available on GitHub and are still under development.

The `cmpreg` package includes functions to fit COM-Poisson models using the parametrization proposed in Chapter 1. In this package, we can specify a regression model for both the mean and dispersion parameter, as discussed in Chapter 3. Several methods are provided for fitted model objects.

The `flexcm` package includes functions to fit the COM-Poisson, Gamma-count, discrete Weibull, generalized Poisson, double Poisson and Poisson-Tweedie, but regression models are allowed for the location parameter only. Actually, for the Poisson-Tweedie model, we implemented a wrapper for the `mcglm::mcglm(..., variance = "poisson_tweedie")`.

The codes below illustrate how to use the functions of these packages. All datasets are included in the packages.

```
#----------------------------------------------------------------------
# Install packages from GitHub (devtools is needed)
#----------------------------------------------------------------------
devtools::install_github("jreduardo/flexcm@v0.0.1")
devtools::install_github("jreduardo/cmpreg@v0.1.0")

#----------------------------------------------------------------------
# Analysis of Sitophilus experiment (Section 4.5.1)
#----------------------------------------------------------------------
data(sitophilus, package = "flexcm")

# Fit all flexible models
form <- ninsect ~ extract
mcmp <- flexcm(form, data = sitophilus, model = "compoisson")
mgct <- flexcm(form, data = sitophilus, model = "gammacount")
mdwe <- flexcm(form, data = sitophilus, model = "discreteweibull")
mgpo <- flexcm(form, data = sitophilus, model = "generalizedpoisson")
mdpo <- flexcm(form, data = sitophilus, model = "doublepoisson")
mptw <- flexcm(form, data = sitophilus, model = "poissontweedie")

# Organize in a list
models <- list(mcmp, mgct, mdwe, mgpo, mdpo, mptw)
names(models) <- c("CMP", "GCT", "DWe", "GPo", "DPo", "PTw")
```

[2] https://github.com/jreduardo/cmpreg
[3] https://github.com/jreduardo/flexcm

```
# Methods for objects for the class 'flexcm'
lapply(models, print)
lapply(models, summary)
lapply(models, equitest)
lapply(models, logLik)
lapply(models, AIC)
lapply(models, fitted)

# Prediction intervals
newdata <- unique(sitophilus[, "extract", drop = FALSE])
lapply(models, function(model) {
    predict(model,
            newdata = newdata,
            type = "response",
            interval = "confidence",
            level = 0.95,
            augment_data = TRUE)
})

#------------------------------------------------------------------------
# Analysis of nitrofen experiment (Sections 3.6.3 and 5.6.1)
#------------------------------------------------------------------------
data(nitrofen, package = "cmpreg")

# Linear predictors
po <- function(x, degree) c(poly(x, degree, raw = FALSE)[, degree])
form0 <-  ~ 1
form1 <-  ~ po(dose, 1)
form2 <-  ~ po(dose, 1) + po(dose, 2)
form3 <-  ~ po(dose, 1) + po(dose, 2) + po(dose, 3)

# Organize in a list
formulas <- list(form0, form1, form2, form3)
names(formulas) <- c("Constant", "Linear", "Quadratic", "Cubic")

# Fit CMP models using joint strategy
modelscmp <-
    lapply(formulas, function(form) {
        cmp(novos ~ dose + I(dose^2) + I(dose^3),
            dformula = form,
            data = nitrofen)
    })
```

```
# Fit double Poisson models (gamlss and gamlss.dist are needed)
modelsdpo <-
    map(formulas, function(form) {
        gamlss::gamlss(novos ~ dose + I(dose^2) + I(dose^3),
                       sigma.formula = form,
                       family = gamlss.dist::DPO,
                       data = nitrofen,
                       trace = FALSE)
    })

# Fit double Generalized Linear Poisson model (dglm is needed)
modelsglm <-
    map(formulas, function(form) {
        dglm::dglm(novos ~ dose + I(dose^2) + I(dose^3),
                   dformula = form,
                   family = poisson,
                   data = nitrofen)
    })

# Methods for objects of class 'cmpreg'
lapply(modelscmp, summary)
lapply(modelscmp, logLik)
lrtest(modelscmp) # or anova(modelscmp)

# Some type of residuals
model <- modelscmp[[2]]
plot(fitted(model), residuals(model))
plot(fitted(model), residuals(model, type = "pearson"))

# Prediction with confidence intervals
newdf <- nitrofen[c(1, 11, 21, 31), -2, drop = FALSE]
predict(object = model,
        newdata = newdf,
        what = "all",
        type = "response")
```

BIBLIOGRAPHY

Aitkin, M. (1987). “Modelling Variance Heterogeneity in Normal Regression Using GLIM”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36.3, pp. 332–339.

Andersen, D. A. and W. H. Bonat (2017). “Double Generalized Linear Compound Poisson models to Insurance Claims Data”. In: *Electronic Journal of Applied Statistical Analysis* 10.2.

Bailer, A. and J. Oris (1994). “Assessing toxicity of pollutants in aquatic systems”. In: *In Case Studies in Biometry*, pp. 25–40.

Bonat, W. H. (2018). “Multiple Response Variables Regression Models in R: The mcglm Package”. In: *Journal of Statistical Software* 84.4, pp. 1–30.

Bonat, W. H. and B. Jørgensen (2016). “Multivariate covariance generalized linear models”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*.

Bonat, W. H., B. Jørgensen, C. C. Kokonendji, and J. Hinde (2018). “Extended Poisson-Tweedie: properties and regression model for count data”. In: *Statistical Modelling* 18.1, pp. 24–49.

Bonat, W. H., W. M. Zeviani, and E. E. Ribeiro Jr (2017). *Regression Models for Count Data: beyond Poisson model*. Goiás, Brazil: XV EMR - Brazilian Regression Model School.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Mächler, and B. M. Bolker (2017). “glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling”. In: *The R Journal* 9.2, pp. 378–400.

Cameron, A. C. and P. K. Trivedi (2013). *Regression Analysis of Count Data*. 2nd edition. Econometric Society Monographs. New York: Cambridge University press.

Chatla, S. B. and G. Shmueli (2018). “Efficient estimation of COM-Poisson regression and a generalized additive model”. In: *Computational Statistics & Data Analysis* 121, pp. 71–89.

Consul, P. C. and F. Famoye (1992). “Generalized Poisson Regression Model”. In: *Communication in Statistics – Theory and Methods* 21.1, pp. 89–109.

Consul, P. C. and G. C. Jain (1973). “A Generalization of the Poisson Distribution”. In: *Technometrics* 15.4, pp. 791–799.

Cox, D. R. (1962). *Renewal Theory*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Cox, D. R. and N. Reid (1987). “Orthogonality and Approximate Conditional Inference (with discussion)”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 49.1, pp. 1–39.

Daly, F. and R. Gaunt (2016). “The Conway-Maxwell-Poisson distribution: Distributional theory and approximation”. In: *ALEA, Latin American Journal of Probability and Mathematical Statistics* 13, pp. 635–658.

Del Castillo, J. and M. Pérez-Casany (1998). “Weighted Poisson Distributions for Overdispersion and Underdispersion Situations”. In: *Annals of the Institute of Statistical Mathematics* 50.3, pp. 567–585.

Demétrio, C. G. B., J. Hinde, and R. A. Moral (2014). “Models for overdispersed data in entomology”. In: *Ecological modelling applied to entomology.* Springer, pp. 219–259.

Efron, B. (1986). “Double Exponential Families and Their Use in Generalized Linear Regression”. In: *Journal of the American Statistical Association* 84.395, pp. 709–721.

Gaunt, R., S. Iyengar, A. Olde Daalhuis, and B. Simsek (2017). “An asymptotic expansion for the normalizing constant of the Conway-Maxwell-Poisson distribution”. In: *Annals of the Institute of Statistical Mathematics* to appear.

Gilbert, P. and R. Varadhan (2016). *numDeriv: Accurate Numerical Derivatives.* R package version 2016.8-1.

Hilbe, J. M. (2014). *Modeling Count Data.* New York: Cambridge University press.

Hinde, J. and C. G. B. Demétrio (1998). “Overdispersion: models and estimation”. In: *Computational Statistics & Data Analysis* 27.2, pp. 151–170.

Huang, A. (2017). “Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts”. In: *Statistical Modelling* 17.6, pp. 1–22.

Johnson, N. L., A. W. Kemp, and S. Kotz (2005). *Univariate Discrete Distributions.* 3rd edition. Series in Probability and Statistics. New Jersey: John Wiley & Sons.

Jørgensen, B. (1997). *The Theory of Dispersion Models.* Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Kanashiro, S., K. Minami, T. Jocys, C. T. dos Santos Dias, and A. R. Tavares (2008). “Alternative substrates to fern tree fiber in the production of ornamental bromeliad”. In: *Pesquisa Agropecuária Brasileira* 43.10.

Klakattawi, H. S., V. Vinciotti, and K. Yu (2018). “A Simple and Adaptive Dispersion Regression Model for Count Data”. In: *Entropy* 20.142.

Kokonendji, C. C. (2014). “Over- and Underdispersion Models”. In: ed. by N. Balakrishnan. John Wiley & Sons. Chap. 30, pp. 506–526.

Lee, Y. and J. A. Nelder (2006). “Double hierarchical generalized linear models (with discussion)”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 55, pp. 139–185.

Lindsey, J. K. (1996). *Parametric Statistical Inference.* New York: Oxford University Press.

Lord, D., S. R. Geedipally, and S. D. Guikema (2010). “Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion”. In: *Risk Analysis* 30.8, pp. 1268–1276.

Lord, D., S. D. Guikema, and S. R. Geedipally (2008). “Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes”. In: *Accident Analysis and Prevention* 40, pp. 1123–1134.

Luyts, M., G. Molenberghs, G. Verbeke, K. Matthijs, E. E. Ribeiro Jr, C. G. B. Demétrio, and J. Hinde (2018). “A Weibull-count approach for handling under- and overdispersed longitudinal/clustered data structures”. In: *Statistical Modelling* to appear.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models.* 2nd edition. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Series in Statistics. New York: Springer.

Molenberghs, G., G. Verbeke, and C. G. B. Demétrio (2007). "An extended random-effects approach to modelling repetead, overdispersed count data". In: *Lifetime Data Analysis* 13, pp. 513–531.

Nakagawa, T. and S. Osaki (1975). "The Discrete Weibull Distribution". In: *IEEE Transactions on Reliability* 24.5, pp. 300–301.

Nelder, J. A. and D. Pregibon (1987). "An Extended Quasi-likelihood Function". In: *Biometrika* 74.2, pp. 221–232.

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135, pp. 370–384.

Nocedal, J. and S. J. Wright (2006). *Numerical optimization*. 2nd edition. Series in Operations Research. New York: Springer, p. 636. ISBN: 0387987932.

Paula, G. A. (2013). "On diagnostics in double generalized linear models". In: *Computational Statistics & Data Analysis* 68, pp. 44–51.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. New York: Oxford University Press.

Pregibon, D. (1984). "Review: P. McCullagh, J. A. Nelder, Generalized Linear Models"". In: *The Annals of Statistics* 12.4, pp. 1589–1596.

Puig, P. and J. Valero (2006). "Count data distributions: some characterizations with applications". In: *Journal of the American Statistical Association* 101.473, pp. 332–340.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ribeiro Jr, E. E., W. M. Zeviani, W. H. Bonat, C. G. B. Demétrio, and J. Hinde (2018). "Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data". In: *arXiv (Statistics Applications and Statistics Methodology)*.

Ribeiro, L. P., J. D. Vendramim, K. U. Bicalho, M. S. Andrade, J. B. Fernandes, R. A. Moral, and C. G. B. Demétrio (2013). "*Annona mucosa* Jacq. (Annonaceae): A promising source of bioactive compounds against *Sitophilus zeamais* Mots. (Coleoptera: Curculionidae)". In: *Journal of Stored Products Research* 55, pp. 6–14.

Rigby, R. A. and D. M. Stasinopoulos (2005). "Generalized additive models for location, scale and shape (with discussion)". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.3, pp. 507–554.

Ross, G. J. S. (1970). "The Efficient Use of Function Minimization in Non-Linear Maximum-Likelihood Estimation". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 19.3, pp. 205–221.

Salvador, E. (2008). "Evaluation of Alternative Substrates to the Xaxim dust for 'Matrogrossense' fern (*Poly Aureum*) Cultivation". In: *Acta Horticulturae* 779, pp. 547–554.

Sellers, K. F. and D. S. Morris (2017). "Underdispersion models: Models that are "under the radar"". In: *Communication in Statistics – Theory and Methods* 46.24, pp. 12075–12086.

Sellers, K. F. and G. Shmueli (2010). "A flexible regression model for count data". In: *Annals of Applied Statistics* 4.2, pp. 943–961.

Serafim, M. E., F. B. Ono, W. M. Zeviani, J. O. Novelino, and J. V. Silva (2012). "Umidade do solo e doses de potássio na cultura da soja". In: *Revista Ciência Agronômica* 43.2, pp. 222–227.

Shmueli, G., T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright (2005). "A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54.1, pp. 127–142.

Silva, A. M., P. E. Degrande, R. Suekane, M. G. Fernandes, and W. M. Zeviani (2012). "Impacto de diferentes níveis de desfolha artificial nos estágios fenológicos do algodoeiro". In: *Revista de Ciências Agrárias* 35.1, pp. 163–172.

Smyth, G. K. (1988). "Generalized Linear Models with Varying Dispersion". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 51.1, pp. 47–60.

Smyth, G. K. and A. P. Verbyla (1999). "Adjusted likelihood methods for modelling dispersion in generalized linear models". In: *Environmetrics* 10.6, pp. 695–709.

Steutel, F. W. and J. G. F. Thiemann (1989). "The gamma process and the Poisson distribution". In: *(Memorandum COSOR; Vol. 8924). Eindhoven: Technische Universiteit Eindhoven.*

Turner, H. and D. Firth (2007). *Generalized nonlinear models in R: An overview of the gnm package.* Tech. rep. ESRC National Centre for Research Methods.

Vieira, A. M., R. A. Leandro, C. G. Demétrio, and G. Molenberghs (2011). "Double generalized linear model for tissue culture proportion data: a Bayesian perspective". In: *Journal of Applied Statistics* 38.8, pp. 1717–1731.

Wedderburn, R. W. M. (1974). "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method". In: *Biometrika* 61.3, p. 439.

Winkelmann, R. (1995). "Duration Dependence and Dispersion in Count-Data Models". In: *Journal of Business & Economic Statistics* 13.4, pp. 467–474.

Winkelmann, R. (2008). *Econometric Analysis of Count Data.* 5th edition. Berlin, Heidelberg: Springer-Velag, p. 342.

Winkelmann, R. and K. F. Zimmermann (1994). "Count Data Models for Demographic Data". In: *Mathematical Population Studies* 4.3, pp. 205–221.

Zamani, H. and N. Ismail (2012). "Functional Form for the Generalized Poisson Regression Model". In: *Communication in Statistics – Theory and Methods* 41, pp. 3666–3675.

Zeviani, W. M., P. J. Ribeiro Jr, W. H. Bonat, S. E. Shimakura, and J. A. Muniz (2014). "The Gamma-count distribution in the analysis of experimental underdispersed data". In: *Journal of Applied Statistics* 41.12, pp. 2616–2626.

Zou, Y., S. R. Geedipally, and D. Lord (2013). "Evaluating the double Poisson generalized linear model". In: *Accident Analysis and Prevention* 59, pp. 497–505.