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

Models for overdispersed, correlated count entomological data

Sidcleide Barbosa de Sousa

Thesis presented to obtanin the degree of Doctor in Science. Área: Statistics and Agricultural Experimentation

Piracicaba
2023

Sidcleide Barbosa de Sousa
Bachelor in Statistics

Models for overdispersed, correlated count entomological data

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

Thesis presented to obtanin the degree of Doctor in Science. Área: Statistics and Agricultural Experimentation

Piracicaba
2023

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

Sousa, Sidcleide Barbosa de

Models for overdispersed, correlated count entomological data / Sidcleide Barbosa de Sousa. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2023 .

86 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Modelos lineares generalizados mistos 2. Efeitos aleatórios 3. Modelos Combinados 4. Dispersão extra . I. Título.

AGKNOWLEDGMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis, especially my parents Severino Lourenço de Sousa and Maria Rosalia Barbosa de Sousa, my sister Sivoneide Barbosa de Sousa, my four brothers, my husband Francisco and my daughter Maria Fernanda for their love and supporting me throughout my life.

To my adviser, Prof. Dr. Clarice Garcia Borges Demétrio, for the continuous support of my Doctorate, for her patience, motivation, enthusiasm and knowledge.

I would like to thank the entire Graduate Program in Statistics and Agricultural Experimentation (PPGEEA) of the Department of Exact Sciences at ESALQ USP. All the teachers and professionals who work in the department and who helped me during this period and all the my friends.

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoas de Nível Superior), Brazil.

SUMMARY

RESUMO

Modelos para dados entomológicos superdispersos, correlacionados na forma de contagens

Abstract

Resultados de interesse na área entomológica estão frequentemente na forma de contagens e como uma primeiro passo, o modelo padrão para análise desse tipo de dados é o modelo de Poisson, um caso particular de modelos lineares generalizados. As suposições básicas para esse modelo são independência das observações e taxa constante de ocorrência dos eventos. Se uma ou ambas suposições falham a variância observada dos dados será maior (menor) do que a variância esperada pelo modelo de Poisson, resultando no que é chamado superdispersão (subdispersão). Muitos modelos diferentes para superdispersão (subdispersão) podem aparecer de mecanismos específicos alternativos para o processo gerador dos dados. Outra razão para estender o modelo de Poisson é devido à ocorrência de uma estrutura hierárquica nos dados resultante de medidas repetidas feitas na mesma unidade experimental. Nas aplicações entomológicas envolvendo dados de contagem, frequentemente, ocorre um excesso de zeros. Neste trabalho, é apresentada uma revisão de modelos que podem ser usados para levar em conta os diversos aspectos de falhas das suposições do modelo Poisson. A metodologia proposta é ilustrada, usando dados de um experimento para avaliar 25 isolados de fungos entomopatogênicos (Metarhizium spp., B. bassiana and I. fumosorosea) e comparar com três tratamentos de referência no controle de *T. urticae*. Comparam-se os resultados e, também, são discutidos aspectos de seleção de modelos e diagnósticos. Para agrupamento dos isolados são propostos dois métodos. todos os métodos foram implementados usando o software `R`.

Abstract
Palavras-chave: Modelos lineares generalizados mistos; Efeitos aleatórios; Modelos Combinados; Dispersão extra.

ABSTRACT

Models for overdispersed, correlated count entomological data

Abstract

Outcomes of interest for entomological data are often in the form of counts and as a first step, a standard model to analyse this type of data is the Poisson model, an example of generalized linear models. The basic model assumptions are independence of observations and constant rate of event occurrence. If one or both of these assumptions failure the variance of the data will be greater (smaller) than the variance expected using the Poisson model resulting in what is called overdispersion (undersispersion). Many different models for overdispersion (underdispersion) can arise from alternative possible mechanisms for the underlying process. Another reason for extending the Poisson model is because of the occurrence of a hierarchical structure in the data caused by a clustering resulted from repeatedly measuring the outcome on the same experimental unit. In entomological applications involving count data there is often an excess of zero observations. In this work we present a review of models that can be used to take into account the different aspects of the failure of the Poisson model assumptions. The proposed methodology is illustrated using data of an experiment to evaluate 25 isolates of entomopathogenic fungi (Metarhizium spp., B. bassiana and I. fumosorosea) and compare with the three reference treatments on the control of *T. urticae*. We compared the results and also discussed model selection and diagnostics. For grouping the isolates we proposed two different methods. All the methods were implemented in the software R.

Abstract
Keywords: Generalized linear mixed model; Random effect; Combined model; Extra-dispersion.

1 MODELS FOR OVERDISPERSED, CORRELATED COUNT ENTOMOLOGICAL DATA

1.1 Introduction

Strawberry is an economically important crop. The largest world strawberry producers are United States, Spain, Japan, Italy, South Korea, and Poland. Spain and the United States are the world's largest strawberry exporters (Sjulin, 2003).

The strawberry stands out among the group of climate fruits, in Brazil where the temperature varies regularly throughout the year, with the average above 10ºC, in the warmer months and between -3ºC and 18ºC in the cold months. The interest in strawberry cultivation is justified by the high profitability of the crop, the wide knowledge and acceptance of the fruit by the consumer, and the diversity of marketing and processing of the strawberry (sweets, yogurt, jellies, juices, pulp, and ice cream). With a production of approximately 105000 tons spread over 4000 hectares, the cultivation is concentrated in the states of Minas Gerais (41.4%), Rio Grande do Sul (25.6%), São Paulo (15.4%), Paraná (4.7%) and Distrito Federal (4%). (Ceuppens et al., 2015).

The occurrence of the main pests of strawberry crop will depend on the region of cultivation, climate, crop treatment and crop management (Kovaleski et al., 2006). The damage is linked to the destruction of the aerial parts of the plant, attack on the fruit and the transmission of viruses that may reduce plant production (Canassa et al., 2020).

Brazil is the world's largest consumer of agrochemicals, and the same may remain in strawberry fruits since, during production, harvests are performed twice a week and the crop receives weekly applications of the products. This may be one of the reasons for which this agricultural product is on the list of foods with high levels of chemical residues annually, endangering the health of humans, as well as causing environmental contamination (ANVISA, 2013).

Effective strategic studies in the control of pests and diseases are necessary and, at the same time, capable of increasing production, with minimal environmental impact. Because of this, the demand for products from organic systems has increased (Castro, 2011). For this reason promising entomopathogenic fungi of *Metarhizium spp.*, *B. bassiana*, *I. fumosorosea* were studied. They are effective in controlling pests, diseases and at the same time being able to promote plant growth, having a high contribution to strawberry crop (Canassa et al., 2020). The aim of many studies is to select isolates that are highly potent.

The class of generalised linear models (GLM) was introduced by Nelder e Wedderburn (1972), for handling a range of statistical models for Gaussian and non-Gaussian data. Outcomes of interest for entomological data are often in the form of counts and as a first step, a standard model to analyse this type of data is the Poisson model, an example

of generalized linear models (McCullagh e Nelder, 1989). The basic model assumptions are independence of observations and constant rate of event occurrence. If one or both of these assumptions failure the variance of the data will be greater than the variance expected using the Poisson model resulting in what is called overdispersion.

There are many different possible causes of overdispersion and in specific situations a number of these could be involved. Some common possibilities in entomological studies are variability of experimental material, correlation between individual responses, cluster and multistage sampling, aggregation, omitted unobserved variables. In general, it is difficult to infer the precise cause, leading to the overdispersion (Demétrio et al., 2014). A number of different models and associated estimation methods have been proposed to take account of overdispersion in order to avoid incorrect inferences (Hinde e Demétrio, 1998).

Many different specific models for overdispersion can arise from alternative possible mechanisms for the underlying process. The simplest way is to assume some more general form for the variance function, possibly including additional parameters, leading to the quasi-poisson model. Another way is to assume a two-stage model for the response, that is, to assume that the basic response model parameter itself has some distribution having as a typical example the negative binomial model. An alternative model arises from the inclusion of random effects in the linear predictor of the model as the Poisson-normal model an example of a generalised linear mixed model (GLMM), allowing to get a measure of intraclass correlation.

Another reason for extending the Poisson model is because of the occurrence of a hierarchical structure in the data caused by a clustering resulted from repeatedly measuring the outcome on the same experimental unit (Verbeeke e Molenberghs, 2000). The possible correlation between measurements for the same individual is often accommodated through the inclusion of subject-specific, random effects. Additionally, overdispersion and correlation between observations may occur simultaneously, and models accommodating them at once are less than common. Molenberghs et al. (2007), and Molenberghs et al. (2010) propose a generalized linear model, accommodating overdispersion and clustering through two separate sets of random effects, of gamma and normal type, respectively. Additionally, one frequent manifestation of overdispersion is that the incidence of zero counts is greater than expected for the Poisson distribution and this is of interest because zero counts frequently have special status (Ridout et al., 1998).

In entomological applications involving count data there is often an excess of zero observations. Poisson regression models provide a standard framework for the analysis of count data but it is not adequate when the incidence of zero counts is greater than expected for the Poisson distribution and this is of interest because zero counts frequently have special status (Ridout et al., 1998). There are two types of zeros that can occur: *structural zeros*, which are inevitable, and *sampling zeros*, which occur by chance. The

distinction between them will depend on the generating process of the count data. It is also possible to have fewer zero count than expected (zero-deflation).

In this work we review and compare methods for analysing count data with particular focus on potential applications in agricultural research. Section 1.2 provides a motivation data set. Section 1.3 presents some models used for the analysis of count data, discusses model selection and diagnostics and gives methods for grouping the isolates. The motivation data set is analysed in Sections 1.4, 1.5 and 1.6. Some general considerations are presented in Section 1.7. The scripts developed in the software **R** (R Core Team, 2020) are presented in the Appendix.

1.2 Case study - description

An experiment in a randomized block complete design with 28 isolates in 5 blocks was conducted in a greenhouse for 180 days at ± 28°C and natural light, with biweekly fertilization. This experiment was repeated twice.

For the first experiment (from July 2016 to January 2017), strawberry plants of cultivar 'Albion' were obtained at 2-4 leaves stage from the seedling nursery "Irmãos Baptistella", Itatiba, São Paulo, Brazil. The aim was to evaluate 25 isolates of entomopathogenic fungi (*Metarhizium spp.*, *B. bassiana* and *I. fumosorosea*) and compare with the three reference treatments *T. harzianum ESALQ 1306*, Quartzo and Control (0.05% Tween 80). Roots of individual strawberry plant were immersed for two min in 30 ml for each treatment. Plants were then directly transplanted individually into 2 L pots containing 50% of surface soil 40% of substrate Tropstrato V-9 Mix and 10% of medium texture sand. The remains of each treatment after root dipping were poured over the soil substrate of the strawberry plant. Sixty days after inoculation of strawberry roots, one *T. urticae* female from the laboratory rearing was placed on a leaflet of each strawberry plant per treatment. After infestation, the leaflet with one *T. urticae* female was covered with a clip cage (4.5 cm high, 3.8 cm diameter) with fine mesh at the open top end (0.09 mm mesh size) preventing the spread of *T. urticae* to other parts of the plant (Figure 1.1). After seven days, each infested leaflet was detached and the number of eggs under the clip cage was counted under a 10X stereoscopic binocular microscope.

For the second experiment (from January to July 2018), the strawberry cultivar was "Pircinque" with seedlings at the 2-4 leaves stage obtained from the seedling nursery "Irmãos Baptistella". The aim was to evaluate the effect of the same 25 isolates, previously used, and compare with the three reference treatments *T. harzianum*, Quartzo and Control, on the number of *T. urticae* eggs at 60 days and 120 days after root inoculation.

Additionally, in both experiments, beneficial effects were evaluated on strawberry plants inoculated with different isolates by counting the total number of leaves per plant at 0, 30, 60, 90, 120, 150 and 180 days after inoculation (DAI), and the total number of

flower per plant at 30, 60, 90, 120, 150 and 180 days after inoculation.

Figure 1.1. Cage clip trap

The exploratory plot of the data (Figure 1.2 (A)) shows that most dispersion plot and (Figure 1.2 (B)) shows the sample variances are greater than the sample means of all experiment, indicating a strong evidence of overdispersion. For experiment I, some treatments show evidence of underdispersion.

1.3 Statistical models

1.3.1 Introduction to Generalized Linear Models (GLMs)

The class of GLMs was introduced by Nelder e Wedderburn (1972) as a framework for handling a range of common statistical models for the analysis of Gaussian and non-Gaussian data. The GLMs are applicable when we have a single response variable Y and p associated explanatory variables. They are defined for three components as follows (Demétrio et al., 2014).

The first component is a set of independent random variables, $Y_1, \ldots, Y_n$, with Y_i following a distribution which is a member of the exponential family distribution

$$f(y_i|\theta_i, \phi) \equiv \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\} \qquad i = 1, \ldots, n \tag{1.1}$$

where ϕ is called the dispersion parameter and θ_i is the canonical parameter, $b(\cdot)$ and $c(\cdot)$ are known functions. Several distributions belongs to the exponential family, e.g. Binomial, Poisson, Normal, Gamma and, Inverse Gaussian. The mean and variance of Y_i

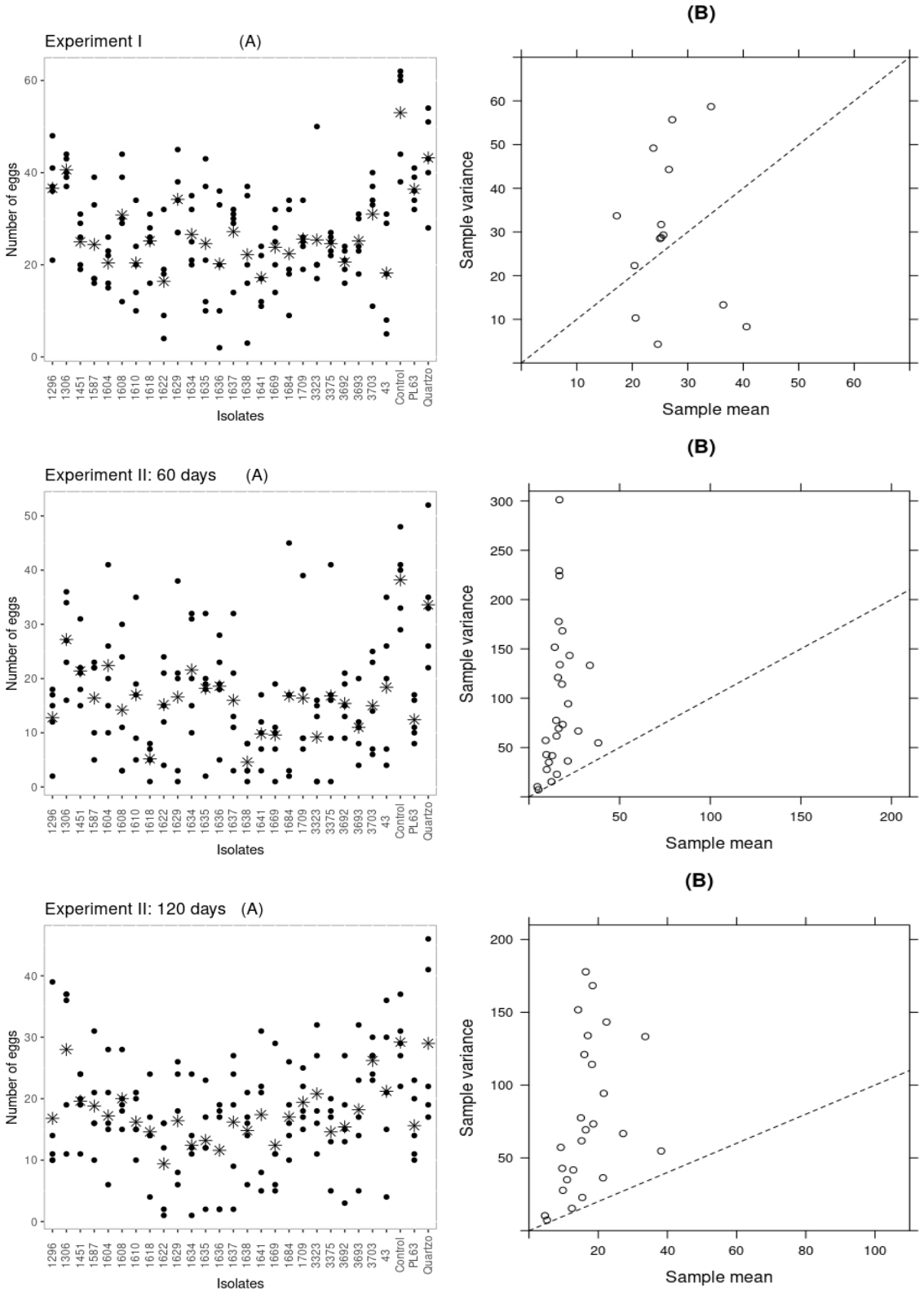

Figure 1.2. (A) Dispersion plots and (B) dispersion plots with sample variance against sample mean of all experiments (dotted line is the identity line).

are given by

$$\mathrm{E}(Y_i) \;=\; \mu_i = \mathrm{b}'(\theta_i)$$

and

$$\mathrm{Var}(Y_i) \quad = \quad \phi \mathrm{b}''(\theta_i) = \phi \mathrm{b}''[\mathrm{b}'(\mu_i)]^{-1} = \phi \mathrm{V}(\mu_i)$$

where $\mathrm{V}(\cdot)$ is called variance function.

For the normal distribution, for example, the mean and variance of the exponential family distributions are related through $\theta_i = \mu_i = \mathrm{b}'(\theta_i)$ and $\phi = \sigma^2$.

The second component, called linear predictor, incorporates into the model the information related to the explanatory variables.

$$\eta_i \quad = \quad \boldsymbol{\beta}' \boldsymbol{x}_i$$

where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $\boldsymbol{x}_i' = [x_{i1}, \ldots, x_{ip}]$ is the i-th row of the $n \times p$ design matrix, $i = 1, \ldots, n$.

The third component, called link function, $g(\cdot)$, provides the relationship between the linear predictor and the mean of the distribution as

$$\eta_i = g(\mu_i) = g(\boldsymbol{x}_i' \boldsymbol{\beta}),$$

where $g(\cdot)$ is a differentiable function.

For a standard generalized linear model maximum likelihood estimates of the regression parameters $\boldsymbol{\beta}$ are easily obtained using an iterative procedure based on a Newton-Raphson or Fisher scoring algorithm.

The analysis of deviance was proposed by Nelder e Wedderburn (1972) to assess the significance of effects in the predictor as a measure that compares a fitted model to the saturated model, and for known ϕ, can be used as a measurement of goodness-of-fit for the fitted model.

The alternative measure of overall fit, the Pearson X^2 statistic, is given by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var}(Y_i)}.$$

1.3.2 Poisson model

The Poisson distribution, a member of the exponential family, is a starting point for the analysis of count data observed over identical time periods. The simplest model assumes that the count random variables Y_i, $i = 1, \ldots, n$, are Poisson distributed with means μ_i, that is, $Y_i \sim \mathrm{Poisson}(\mu_i)$. The probability function can be written as

$$P(Y_i = y_i) \quad = \quad \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots, \quad \mu_i > 0, \tag{1.2}$$

with log-link and linear predictor

$$\log(\lambda_i) = \boldsymbol{x}_i' \boldsymbol{\beta}.$$

The mean and variance of the Poisson model are $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \mu_i$. This implies an index of dispersion $\frac{\mathrm{Var}(Y_i)}{\mathrm{E}(Y_i)} = 1$, a very restrictive assumption when comparing the sample average with the sample variance for a particular set of data.

The deviance for the Poisson model is given by

$$D_p = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

where $\hat{\mu}_i$, $i = 1, 2, \ldots, n$ are the fitted values for the current model. The deviance D_P can be viewed as a measure of goodness-of-fit of the fitted model with p estimated parameters.

The Pearson X^2 statistic, takes the familiar form

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var}(Y_i)} = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

For large expected counts μ_i, D_P and X_P^2 are equivalent and asymptotically both have an approximate X^2 distribution with $n - p$ degrees of freedom.

Then, for a well-fitting model we would expect that D_P and X_P^2 would be approximately equal to the residual degrees of freedom. When this does not happen one explanation is that the variation may simply be different from that predicted by the model (Hinde e Demétrio, 1998; Demétrio et al., 2014). When the variability of the data is smaller (underdispersion) or greater (overdispersion) than the mean, the Poisson model does not fit to the data.

While he phenomenon of overdispersion is well known in literature, underdispersion is less reported. Overdispersion may occur due to the absence of relevant covariates, heterogeneity of sampling units and excess of zeros (Demétrio et al., 2014) and it is important to have models that take into account these features in order to avoid incorrect inferences (Hinde e Demétrio, 1998). Therefore, extensions of the Poisson model can be used to analyze underdispersed or overdispersed data.

Several models were proposed for the analysis of overdispersed count data, including Breslow (1984) and Lawless (1987) and more general discussions are also to be found in McCullagh e Nelder (1989) and Lindsey (1995). We will begin by considering a quasi-likelihood approach to accommodate increased variability.

1.3.3 Quasi-Poisson model

The simplest way of modeling overdispersion is to replace the variance function of the original model by the more general form

$$\mathrm{Var}(Y_i) = \phi\mu_i \tag{1.3}$$

where ϕ is called the dispersion parameter (called heterogeneity factor). A quasi-likelihood method, which requires the specification of the first and second moments of the distribution, is used for estimating $\boldsymbol{\beta}$ and the additional parameter ϕ. The overdispersion

parameter $\phi > 1$ is considered as an unknown, indicates that the increased variation for observation Y_i does not depend on the mean μ_i.

According to Wedderburn (1974), the estimates of the regression parameters β using maximum quasi-likelihood for this constant overdispersion model are identical to those from the Poisson model. However, the assumed greater variability in (1.3) inflates the standard errors of $\hat{\beta}$ by a factor of $\sqrt{\phi}$ compared to those of the Poisson ($\phi = 1$) model. For the quasi-Poisson model (1.3) the estimate of ϕ is

$$\tilde{\phi} = \frac{X_P^2}{n-p}$$

where $X_P^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/\hat{\mu}_i$ is the generalised Pearson statistic for the Poisson model, a measure of goodness-of-fit, and use $\tilde{\phi}$ estimated value to obtain the standard errors of $\hat{\beta}$.

1.3.4 Negative Binomial model

An alternative approach to account for overdispersion in count data is through a two-stage model. Assuming that the conditional distribution of Y_i given z_i is a Poisson model, that is, $Y_i|z_i \sim \text{Poisson}(z_i)$, and that Z_i is a random variable with no particular distributional form with $\text{E}(Z_i) = \mu_i$ and $\text{Var}(Z_i) = \sigma_i^2$. The marginal mean and variance are given by

$$\text{E}[\text{E}(Y_i|z_i)] = \mu_i,$$

and

$$\text{Var}(Y_i) = \text{E}[\text{Var}(Y_i|z_i)] + \text{Var}[\text{E}(Y_i|\lambda_i)] = \mu_i + \sigma_i^2.$$

If we assume $Z_i \sim \text{Gamma}(\alpha, \theta_i)$, a natural and flexible family of distributions on $(0, \infty)$ with a fixed shape parameter α and a varying scale parameter θ_i, that is, with a density

$$f(z_i) = \frac{1}{\theta_i^{\alpha}\Gamma(\alpha)} z_i^{\alpha-1} \text{e}^{-\frac{z_i}{\theta_i}},$$

where $\Gamma(\cdot)$ is the gamma function, $\text{E}(Z_i) = \mu_i = \alpha\theta_i$ and $\text{Var}(Z_i) = \sigma_i^2 = \alpha\theta_i^2$, then unconditionally Y_i has a negative binomial distribution with probability function

$$\begin{aligned} P(Y_i = y_i) &= \frac{1}{\theta_i^{\alpha}\Gamma(\alpha)} \int_0^{\infty} \frac{z_i^{y_i}\text{e}^{-z_i}}{y_i!} z_i^{\alpha-1}\text{e}^{-\frac{z_i}{\theta_i}} dz_i \\ &= \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)\theta_i^{\alpha} y_i!} \left(\frac{\theta_i}{\theta_i+1}\right)^{y_i+\alpha} \\ &= \binom{\alpha+y_i-1}{\alpha-1} \frac{\mu_i^{y_i}\alpha^{\alpha}}{(\mu_i+\alpha)^{y_i+\alpha}}. \end{aligned}$$

with mean

$$\text{E}(Y_i) = \text{E}[\text{E}(Y_i|z_i)] = \text{E}[Z_i] = \alpha\theta_i = \mu_i,$$

and variance

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathrm{E}[\mathrm{Var}(Y_i|Z_i)] + \mathrm{Var}[\mathrm{E}(Y_i|Z_i)] = \mathrm{E}[Z_i] + \mathrm{Var}(Z_i) \\ &= \alpha\theta_i + \alpha\theta_i^2 = \mu_i(1+\theta_i) \end{aligned} \tag{1.4}$$

According to Demétrio et al. (2014), an advantage of using a fixed value of α is that the resulting distribution for Y_i is in the exponential family and so we are still in the generalized linear modelling framework. The negative binomial distribution is similar to the Poisson distribution, but incorporates a variance that is larger than its mean. As a result, it is more flexible and can accommodate more distributional shapes than the Poisson distribution (Gbur et al., 2012).

1.3.5 Poisson-Normal model

Another way to model overdispersion for count data consists in adding an observation level random effect to the linear predictor (Hinde e Demétrio, 1998). Assuming that the conditional distribution of Y_i given z_i is a Poisson model, that is, $Y_i|z_i \sim \mathrm{Poisson}(\lambda_i)$, with log-link and linear predictor

$$\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma z_i$$

where Z_i is a random variable with a standard normal distribution, that is, $Z_i \sim \mathrm{N}(0,1)$. According to Hinde (1982) this additional random effect is a combination of many unexplained things. This is the simpler case of a generalized linear mixed model. There is no closed form for the distribution of Y_i but its mean and variance are given, respectively, by

$$\mathrm{E}(Y_i) = \mathrm{E}[\mathrm{E}(Y_i|z_i)] = \mathrm{E}[\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma Z_i)] = \mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta} + \frac{1}{2}\sigma^2} = \mu_i$$

and

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathrm{E}[\mathrm{Var}(Y_i|z_i)] + \mathrm{Var}[\mathrm{E}(Y_i|z_i)] = \mathrm{E}[\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma Z_i)] + \mathrm{Var}[\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma Z_i)] \\ &= \mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta} + \frac{1}{2}\sigma^2} + \mathrm{e}^{2\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma^2}(\mathrm{e}^{\sigma^2} - 1) = \mu_i + \phi\mu_i^2 \end{aligned} \tag{1.5}$$

So the form of the variance of the Poisson-normal model (1.5) is the same as for the negative binomial distribution (1.4). This implies that approximate quasi-likelihood estimates are the same for both the negative binomial and Poisson-normal models but full maximum likelihood estimates will differ (Hinde e Demétrio, 1998). Hinde (1982) gives the details of maximum likelihood estimation for the Poisson-normal model based on using Guassian-quadrature to integrate over the random effect.

1.3.6 Overdispersed models for longitudinal/correlated data

In many entomological experiments, besides the problem of overdispersion, the studies can be carried out in such a way that several measurements are taken from the same subject or sample unit over time, characterizing a longitudinal study. To analyze this type of data, the univariate models just described can be extended to take into account overdispersion and/or the correlation between the data resulted from repeatedly measuring the outcome on the same experimental unit (Molenberghs et al., 2007, 2017).

To take into account the possible correlation between measurements for the same individual appropriate statistical approaches are needed, such as Generalized Linear Mixed Model (GLMM) and it is often accommodated through the inclusion of subject-specific, random effects as an extension of the linear mixed model in the context of non Gaussian repeated measurements (Verbeeke e Molenberghs, 2000).

Let Y_{ij} be the j-th outcome measured for subject $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ and group the n_i measurements into a vector $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})'$ with a distribution and with a vector of means $\boldsymbol{\lambda_i} = (\lambda_{i1}, ..., \lambda_{in_i})'$.

Repeated-version of the quasi-likelihood model

Here as in Molenberghs et al. (2007), we assume that $Y_{ij}|\lambda_{ij} \sim \text{Poisson}(\lambda_{ij})$ and that $\boldsymbol{\lambda_i} = (\lambda_{i1}, ..., \lambda_{in_i})'$ is a vector of random variables with no particular distributional form with $\text{E}(\boldsymbol{\lambda_i}) = \boldsymbol{\mu}_i$ and $\text{Var}(\boldsymbol{\lambda_i}) = \boldsymbol{\Sigma_i}$. The marginal mean and variance are given by

$$\text{E}[\text{E}(\boldsymbol{Y_i}|\boldsymbol{\lambda_i})] = \boldsymbol{\mu}_i,$$

and

$$\text{Var}(\boldsymbol{Y_i}) = \text{E}[\text{Var}(\boldsymbol{Y_i}|\boldsymbol{\lambda_i})] + \text{Var}[\text{E}(\boldsymbol{Y_i}|\boldsymbol{\lambda_i})] = \boldsymbol{M}_i + \boldsymbol{\Sigma_i},$$

where $\boldsymbol{M_i}$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal. Alternatively, we can assume a gamma distribution for $\boldsymbol{\lambda_i}$, leading to the negative-binomial model.

Repeated-version of the Poisson-normal model

In general, we assume, conditionally on q-dimensional random effects $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$, with density $f(\boldsymbol{b}_i|D)$, the responses Y_{ij} are independent with distributions that are members of exponential family of the form

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) \;\; = \;\; \exp\{\phi^{-1}[y_{ij}\theta_{ij} - b(\theta_{ij})] + c(y_{ij}, \phi)\}, \tag{1.6}$$

with

$$\eta[b'(\theta_{ij})] \;\; = \;\; \eta(\mu_{ij}) = \eta[\text{E}(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta})] = \boldsymbol{x'_{ij}\beta} + \boldsymbol{z'_{ij}b}_i, \tag{1.7}$$

for a known link function $\eta(\cdot)$, with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, β a p-dimensional vector of unknown fixed regression coefficients,

and ϕ a scale parameter. For a count response as proposed by Molenberghs et al. (2007) we have

$$\begin{aligned} Y_i|\boldsymbol{b}_i &\sim \text{Poisson}(\boldsymbol{\lambda_i}), \\ \log(\boldsymbol{\lambda_i}) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_i, \\ \boldsymbol{b}_i &\sim \text{N}(\boldsymbol{0}, D). \end{aligned}$$

The marginal mean vector and variance-covariance matrix of $\boldsymbol{Y}_i$, are given, respectively, by

$$\begin{aligned} \text{E}(\boldsymbol{Y_i}) &= \text{E}[\text{E}(\boldsymbol{Y_i}|\boldsymbol{b}_i)] = \text{E}[\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_i)] \\ &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\text{E}[\exp(\boldsymbol{z}_i'\boldsymbol{b}_i)] = \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_i'\boldsymbol{D}\boldsymbol{z}_i\right) = \boldsymbol{\mu}_i \end{aligned} \tag{1.8}$$

and

$$\begin{aligned} \text{Var}(\boldsymbol{Y_i}) &= \text{E}[\text{Var}(\boldsymbol{Y_i}|\boldsymbol{b}_i)] + \text{Var}[\text{E}(\boldsymbol{Y_i}|\boldsymbol{b}_i)] \\ &= \text{E}[\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_i)] + \text{Var}[\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_i)] \\ &= \boldsymbol{M_i} + \boldsymbol{M_i}\{\exp(\boldsymbol{z}_i'\boldsymbol{D}\boldsymbol{z}_i) - \boldsymbol{J_{n_i}}\}\boldsymbol{M_i}. \end{aligned}$$

where $\boldsymbol{M_i}$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal.

Molenberghs et al. (2007) also derived an expression for the joint probability of $\boldsymbol{Y}_i$. Estimates of $\boldsymbol{\beta}$, $\boldsymbol{D}$ and ϕ for GLMM are obtained from maximizing the marginal likelihood, integrating out the random effects and written as

$$L(\boldsymbol{\beta}, \boldsymbol{D}, \phi) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{b}_i|\boldsymbol{D}) d\boldsymbol{b}_i. \tag{1.9}$$

The problem in maximizing equation (1.9) is the presence of N integrals over the q-dimensional random effects $\boldsymbol{b}_i$. However, the Laplace method works well for a considerable number of mixed models and is implemented in a wide range of software packages as `glmer` and `lme4` in `R`.

Repeated-version of the Poisson-gamma-normal model (Combined Model)

Overdispersion and correlation between observations may occur simultaneously, and models accommodating both at once were proposed by Molenberghs et al. (2007), and Molenberghs et al. (2010) through two separate sets of random effects, of gamma and normal type, respectively. This led to an unified modeling framework, which they termed the combined model. Combining overdispersion and normal random effects, and using the generalized linear model framework, produces the following general family

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \theta_{ij}, \phi) = \exp\{\phi^{-1}[y_{ij}\lambda_{ij} - b(\lambda_{ij})] + c(y_{ij}, \phi)\}, \tag{1.10}$$

with notation similar to the one used in equation (1.6), but with the conditional mean as

$$\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \theta_{ij}) = b'(\lambda_{ij}) = \theta_{ij}k_{ij} = \mu_{ij}, \tag{1.11}$$

where the random effect to accommodate overdispersion acts multiplicatively in the mean of the variable while the normal random effect to capture correlation among repeated observations is placed in the linear predictor.

A model for repeated Poisson data with overdispersion can then be expressed by

$$\begin{aligned} Y_i|\boldsymbol{b}_i, \theta_{ij} &\sim \text{Poisson}(\theta_{ij}k_{ij}), \\ k_{ij} &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}_i), \\ \boldsymbol{b}_i &\sim \mathrm{N}(\boldsymbol{0}, D), \\ \theta_{ij} &= \text{Gamma}(\alpha_{ij}, \beta_{ij}) \end{aligned}$$

resulting a Poisson-Gamma-Normal model, having as a special case the Negative Binomial Normal model. Assuming that $\boldsymbol{\theta}_i$ and $\boldsymbol{b}_i$ are independent and given that $\mathrm{E}(\theta_{ij}) = \alpha_{ij}\beta_{ij}$ and $\mathrm{Var}(\theta_{ij}) = \alpha_{ij}\beta_{ij}^2$, and normal random effects $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$, the marginal mean and variance of Y_{ij} are given, respectively, by

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \mathrm{E}\{\mathrm{E}[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \theta_{ij})]\} = \mathrm{E}\{\mathrm{E}[\theta_{ij}\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i)]\} \\ &= \mathrm{E}[\mathrm{E}(\theta_{ij})\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i)] \quad = \quad \mathrm{E}(\theta_{ij})\mathrm{E}[\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i)] \\ &= \alpha_{ij}\beta_{ij}\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}'\boldsymbol{D}\boldsymbol{z}_{ij}') = \mu_{ij}, \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(Y_{ij}) &= \mathrm{E}\{\mathrm{E}[\mathrm{Var}(Y_{ij}|\boldsymbol{b}_i, \theta_{ij})]\} + \mathrm{E}\{\mathrm{Var}[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \theta_{ij})]\} + \mathrm{Var}\{\mathrm{E}[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i, \theta_{ij})]\} \\ &= \mathrm{E}[\mathrm{E}(\theta_{ij}k_{ij})] + \mathrm{E}[\mathrm{Var}(\theta_{ij}k_{ij})] + \mathrm{Var}[\mathrm{E}(\theta_{ij}k_{ij})] \\ &= \mathrm{E}[\mathrm{E}(\theta_{ij})k_{ij}] + \mathrm{E}[k_{ij}^2\mathrm{Var}(\theta_{ij})] + \mathrm{Var}[\mathrm{E}(\theta_{ij})k_{ij}] \\ &= \mathrm{E}(\theta_{ij})\mathrm{E}(k_{ij}) + \mathrm{E}\{k_{ij}^2[\mathrm{E}(\theta_{ij}^2) - \mathrm{E}(\theta_{ij})^2]\} + \mathrm{E}(\theta_{ij})^2\mathrm{Var}(k_{ij}) \\ &= \mathrm{E}(\theta_{ij})\mathrm{E}(k_{ij}) + \mathrm{E}(k_{ij}^2)\mathrm{E}(\theta_{ij}^2) - \mathrm{E}(k_{ij}^2)\mathrm{E}(\theta_{ij})^2 + \mathrm{E}(\theta_{ij})^2[\mathrm{E}(k_{ij}^2) - \mathrm{E}(k_{ij})^2] \\ &= \mathrm{E}(\theta_{ij})\mathrm{E}(k_{ij}) + \mathrm{E}(k_{ij}^2)\mathrm{E}(\theta_{ij}^2) - \mathrm{E}(k_{ij}^2)\mathrm{E}(\theta_{ij})^2 + \mathrm{E}(\theta_{ij})^2\mathrm{E}(k_{ij}^2) - \mathrm{E}(\theta_{ij})^2\mathrm{E}(k_{ij})^2 \\ &= \mathrm{E}(\theta_{ij})\mathrm{E}(k_{ij}) + \mathrm{E}(k_{ij}^2)\mathrm{E}(\theta_{ij}^2) - \mathrm{E}(\theta_{ij})^2\mathrm{E}(k_{ij})^2 \\ &= \alpha_{ij}\beta_{ij}\exp\{\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}'\boldsymbol{D}\boldsymbol{z}_{ij}\} + \alpha_{ij}\beta_{ij}^2\exp\{2\boldsymbol{x}_{ij}'\boldsymbol{\beta} + 2\boldsymbol{z}_{ij}'\boldsymbol{D}\boldsymbol{z}_{ij}\} + \\ &+ \alpha_{ij}^2\beta_{ij}^2\exp\{2\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{D}\boldsymbol{z}_{ij}\}[\exp(\boldsymbol{z}_{ij}'\boldsymbol{D}\boldsymbol{z}_{ij}) - 1], \end{aligned} \tag{1.12}$$

Molenberghs et al. (2007) also derived an expression for the joint probability of $\boldsymbol{Y}_i$, and showed that fitting the combined model proceeds by integrating over the random effects. The joint distribution of the ij-th observation, assuming θ_{ij} and $\boldsymbol{b}_i$ are independent, is given by

$$f_i(y_{ij}|\beta, \boldsymbol{b}_i, \theta_{ij}) \quad = \quad f_{ij}(y_{ij}|\beta, \boldsymbol{b}_i, \boldsymbol{\theta}_i)f(\boldsymbol{b}_i|\boldsymbol{D})f(\boldsymbol{\theta}_i|\alpha_i, \beta_i) \tag{1.13}$$

The likelihood contribution of subject i is

$$f_i(y_i|\boldsymbol{\beta}, \boldsymbol{D}, \alpha_i, \beta_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, \boldsymbol{b}_i, \boldsymbol{\theta}_i) f(\boldsymbol{b}_i|\boldsymbol{D}) f(\boldsymbol{\theta}_i|\alpha_i, \beta_i) d\boldsymbol{b}_i d\boldsymbol{\theta}_i. \quad (1.14)$$

where, β groups all parameters in the conditional model for $\boldsymbol{Y}_i$. From equation (1.14) the likelihood derives as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{D}, \alpha, \beta) &= \prod_{i=1}^{N} f_i(y_i|\beta, \boldsymbol{D}, \alpha_i, \beta_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\theta}_i) f(\boldsymbol{b}_i|\boldsymbol{D}) f(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i) d\boldsymbol{b}_i d\boldsymbol{\theta}_i \end{aligned} \quad (1.15)$$

The problem in maximizing equation (1.15) is the presence of N integrals over the random effects $\boldsymbol{b}_i$ and $\boldsymbol{\theta}$. The standard software tools, such as the `glmer` and `lme4` in the `R`, can be used for maximum likelihood estimation.

1.3.7 COM-Poisson model

The COM–Poisson distribution, proposed for Conway e Maxwell (1962), is a two-parameter generalization of the Poisson distribution (Sellers et al., 2010; Ribeiro Jr et al., 2020). A random variable $Y_i \sim$ COM-Poisson(λ_i, ν) has a probability mass function given by

$$P(Y_i = y_i) = \frac{\lambda_i^{y_i}}{(y_i!)^{\nu} Z(\lambda_i, \nu)}, \qquad y_i = 0, 1, 2, \ldots, \qquad i = 1, 2, \ldots, n, \quad (1.16)$$

where $\lambda_i > 0$, $\nu \geq 0$ and

$$Z(\lambda_i, \nu) = \sum_{J=0}^{\infty} \frac{\lambda_i^j}{(j!)^{\nu}},$$

is a normalizing constant that depends on both parameters. This distribution can handle under- ($\nu > 1$), over- ($0 < \nu < 1$) and equidispersion ($\nu = 1$).

There is no closed form for its moments. Shmueli et al. (2005) using an asymptotic approximation for $Z(\lambda, \nu)$, showed that the mean and variance of the COM-Poisson distribution can be approximated by

$$\mathrm{E}(Y_i) \approx \lambda^{\frac{1}{\nu}} - \frac{\nu - 1}{2\nu} \quad \text{and} \quad \mathrm{Var}(Y_i) \approx \frac{\lambda^{\frac{1}{\nu}}}{\nu}, \quad (1.17)$$

with accurate approximations for $\nu \leq 1$ or $\lambda > 10^{\nu}$ (Sellers et al., 2010).

The regression COM-Poisson model was proposed by Sellers et al. (2010) to modelling the relationship between $\mathrm{E}(Y_i)$ and the covariates x_i, indirectly, using the log link function

$$\eta_i = \log(\lambda_i) = x_i' \boldsymbol{\beta}.$$

A reparametrization of the COM-Poisson model, based on the mean approximation given by equation (1.17), was proposed by Ribeiro Jr et al. (2020), introducing a new parameter

$$\mu \;= h_\nu(\lambda) = \lambda^{\frac{1}{\nu}} - \tfrac{\nu-1}{2\nu} \quad \Longrightarrow \quad \lambda = h_\nu^{-1}(\mu) = \left(\mu + \tfrac{(\nu-1)}{2\nu}\right)^{\nu} \tag{1.18}$$

and taking the precision parameter on the log scale, $\phi = \log(\nu)$, to avoid restrictions on the parameter space, $\phi \in \mathbb{R}$.

The reparametrized COM-Poisson$_\mu$ probability mass distribution function is given by

$$P(Y_i = y_i) \;=\; \left(\mu + \frac{e^\phi - 1}{2e^\phi}\right)^{ye^\phi} \frac{(y!)^{-e^\phi}}{Z(\mu,\phi)}, \quad y = 0, 1, 2, \ldots, \tag{1.19}$$

where $\mu > 0$ allowing for modelling overdispersion ($\phi < 0$), underdispersion ($\phi > 0$) and equidispersion ($\phi = 0$) (Ribeiro Jr et al., 2020).

The parameter estimates are obtained by numerical maximization of the log-likelihood function, using `BFGS` algorithm, with the Hessian matrix calculated numerically by finite differences (Richardson method). The inferences are based on standard asymptotic likelihood theory (Ribeiro Jr et al., 2020). The log likelihood function from the COM-Poisson$_\mu$ distribution is given by

$$L(y|\phi,\boldsymbol{\beta}) \;=\; e^\phi \left[\sum_{i=1}^{n} y_i \log\left(\mu_i + \frac{e^\phi - 1}{2e^\phi}\right) - \sum_{i=1}^{n} \log(y_i!)\right] - \sum_{i=1}^{n} \log[Z(\mu_i,\phi)], \tag{1.20}$$

where $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, let $\boldsymbol{x}_i' = (x_1, \ldots, x_n)$. The normalizing constant $Z(\mu_i, \phi)$ is approximated by

$$Z(\mu,\phi) \;=\; \sum_{i=1}^{\infty} \left[\left(\mu_i + \frac{e^\phi - 1}{2e^\phi}\right)^{je^\phi} \frac{1}{(j!)^{e^\phi}}\right]. \tag{1.21}$$

Ribeiro Jr et al. (2019) proposed an extension of the COM-Poisson model to jointly model the mean and the dispersion as functions of covariates taking into account, possibly, under- and overdispersion in the same count data set. Estimation and inference are based on the maximum likelihood method.

Let y_i, $i = 1, 2, \ldots, n$, be independent realizations of Y_i from COM-Poisson distributions with parameters μ_i and ν_i. The proposed COM-Poisson varying dispersion model assumes

$$\begin{aligned} Y_i|\boldsymbol{x}_i &\sim \text{CMP}_\mu(\boldsymbol{\mu_i}, \boldsymbol{\nu_i}), \\ \eta_i = g(\mu_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}, \\ \xi_i = h(\nu_i) &= \boldsymbol{z}_i'\boldsymbol{\gamma}, \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)^T$ are the parameters to be estimated, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{np})^T$ and $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{nq})^T$ are vectors of known covariates, and $g(\cdot)$ and $h(\cdot)$ are suitable link functions, such as the log. Maximum likelihood estimation for fitting reparametrized (and original) version of COM-Poisson models with varying dispersion and methods for computing the associated confidence intervals are implemented in the `R` package `cmpreg` (`https://github.com/jreduardo/cmpreg`).

1.3.8 Zero-Inflated Models

In entomological applications involving count data there is often an excess of zero observations. Poisson regression models provide a standard framework for the analysis of count data but it is not adequate when the incidence of zero counts is greater than expected for the Poisson distribution and this is of interest because zero counts frequently have special status (Ridout et al., 1998). There are two types of zeros that can occur: *structural zeros*, which are inevitable, and *sampling zeros*, which occur by chance. The distinction between them will depend on the generating process of the count data. It is also possible to have fewer zero count than expected (zero-deflation).

Different types of models have been proposed in the literature to take into account overdispersion caused by the zero-inflation. The mixed Poisson distributions have been used widely to model overdispersed data. The most used distribution is the negative binomial, that has a higher probability for zero than the Poisson distribution.

Zero-inflated models are two component mixture models combining excess zero with a count distribution such as Poisson or negative binomial (Ridout et al., 1998). To modify the standard Poisson distribution to allow for extra zeros using a *zero-inflated Poisson* (ZIP) distribution, we augment the probability of zero by a proportion w and for the remainder the Poisson parameter takes the fixed value λ, given by

$$P(Y = y) = \begin{cases} w + (1-w)\exp(-\lambda) & y = 0 \\ (1-w)\frac{\exp(-\lambda)\lambda^y}{y!} & y > 0. \end{cases} \tag{1.22}$$

According to Ridout et al. (1998) it is possible for w in equation (1.22) to assume negative values, giving a *zero–inflated* distribution, Zero-inflated data seldom arise in practice, however, and we shall assume $0 \leq w < 1$. For the zero-inflated Poisson distribution, the mean is

$$\mathrm{E}(Y) = (1-w)\lambda = \mu$$

while the variance is

$$\mathrm{Var}(Y) = \mu + \mu^2 \left(\frac{w}{1-w} \right) \tag{1.23}$$

The variance given by (1.23) has the same form as equations (1.4) and (1.5) but resulted from different generating process. We may think of it as a model for overdispersed count

data, but data in which the overdispersion arises in a very specific way, through an excess of zeros.

In mixed Poisson models, covariates are introduced via a log-linear model for λ, as in the Poisson model and logit model for w

$$\log(\lambda) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{w}{1-w}\right) = \boldsymbol{Z}\boldsymbol{\gamma}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are matrices of covariates and β and γ are vectors of parameters. The two sets of covariates may or may not coincide. For zero-inflated models it is also possible to include random effects in the linear predictor to account for overdispersion and correlation between clustered data.

An alternative model that takes into account zero-inflation and extra overdispersion is a *zero-inflated negative binomial* (ZINB) distribution (Ridout et al., 2001), given by

$$P(Y=y) \quad = \quad \begin{cases} w + (1-w)(1+\alpha\lambda^c)^{-\frac{\lambda^{1-c}}{\alpha}}, & y=0 \\ (1-w)\frac{\Gamma(y+\frac{\lambda^{1-c}}{\alpha})}{y!\Gamma(\frac{\lambda^{1-c}}{\alpha})} \quad (1+\alpha\lambda^c)^{-\frac{\lambda^{1-c}}{c}} \quad (1+\frac{\lambda^{-c}}{\alpha})^{-y} & y>0 \end{cases} \tag{1.24}$$

where $\alpha(\geq 0)$ is a dispersion parameter that is assumed not to depend on covariates. This distribution reduces to the zero-inflated Poisson distribution in the limit $\alpha \to 0$.

The mean of the distribution is

$$\mathrm{E}(Y) = (1-w)\lambda = \mu$$

while the variance of the distribution is

$$\mathrm{Var}(Y) = (1-w)\lambda(1+w\lambda+\alpha\lambda^c).$$

The index c identifies the particular form of the underlying negative binomial distribution with mean λ, for $c=0$, the variance of the negative binomial distribution is $(1+\alpha)\lambda$ (NB1) and, for $c=1$, the variance is $\lambda+\alpha\lambda^2$ (NB2). In the same way as for ZIP models, covariates are introduced via a log-linear model for λ, and logit model for w.

Zero-inflated count data models can be fitted using the `zeroinfl()` function from the `pscl` package, in the software **R** (R Core Team, 2020). It allows the fitting of zero-inflated Poisson and negative binomial models with regression models for both components, but without additional random effects (Zeileis et al., 2008).

1.3.9 Model selection and diagnostics

A model selection process involves a combination of choosing an adequate distribution and link function, testing for terms of possible interest and the use of the goodness-of-fit to check that any selected models are adequate descriptions of the data (Moral et al.,

2017). An approach for checking the goodness of fit of a model is to use the half normal plot with simulated envelope (`hnp`). The steps for building the envelope (Hinde e Demétrio, 1998) are

I. After fitting a model, extract the values of a chosen diagnostic quantity, and get its absolute values, $d_{(i)}$;

II. Perform m simulations, considering the fitted model with the same values for the explanatory variables;

III. Fit the same model to each of the m samples, and get the ordered absolute values of the diagnostic quantity, $q^*_{j(i)}$, $j = 1, \ldots, m$, $i = 1, \ldots, n$;

IV. For each i, calculate the mean, first and third quartiles of the $q^*_{j(i)}$;

V. Plot these values and the observed $d_{(i)}$ versus the half-normal order statistics given by

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right), \tag{1.25}$$

where $\Phi^{-1}(\cdot)$ is an accumulated function of the standard normal distribution, and n refers to the sample size obtained, with $i = 1, \ldots, n$.

This type of graph is implemented through the function `hnp` for some probability distributions, using *software* **R** (Moral et al., 2017). If the model fits to the data we would expect the plot of the observed values to lie within the boundaries of the envelope (Hinde e Demétrio, 1998).

The selection of the linear predictor for a model, in general, involve comparisons of nested models and deviance differences (Analysis of deviance), that is, likelihood ratio tests. Involves evaluating the value of the likelihood function for the complete model and evaluating the value of the likelihood function for the model under the conditions of H_0 (reduced model), using ML. The nested and reference models have the same set of covariance parameters but different sets of fixed-effect parameters

$$\begin{aligned} LR &= -2[\text{logLik(reduced model) - logLik(complete model)}] \\ &= \text{deviance(reduced model) - deviance(complete model)} \end{aligned}$$

where logLik is the logarithm of the likelihood function. $LR \sim \chi^2_\nu$ where ν is the difference in number of fixed-effect parameters between the two models.

Many interesting comparisons involve non-nested models and in this case we can use of the Akaike Information Criterion (AIC; (Akaike, 1973)) or Bayes Information Criterion (BIC; (Schwarz, 1978))

$$\text{AIC} = -2\text{logLik} + 2p$$

and

$$\text{BIC} = -2\text{logLik} + \log(n)p$$

where p is the number of fitted parameters and n is the number of observations.

1.3.10 Clustering methods for the means

Empirical grouping - Visual inspection

As suggested by Fatoretto et al. (2018), the isolates can be grouped and a likelihood-ratio test can be performed to test the differences of the similarity observed between isolates and the groups can be constructed according to the predicted values in order to group the ones that have a similar behaviour. For the quasi-likelihood model the nested models are compared using the F-test.

K-means clustering

Another way to group data sets is to use the K-means algorithm proposed by MacQueen (1967) which has a wide application and according to it each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid).

Given a set of observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, where each observation $\boldsymbol{x}_i$ is a d-dimensional real vector, the aim of K-means clustering is to partition n observations into ($k \leq n$) clusters $\boldsymbol{C} = \{C_1, C_2, \ldots C_k\}$ so as to minimize the within-cluster variances (squared Euclidean distances), that is, minimizes

$$\sum_{j=1}^{k} \sum_{\boldsymbol{x} \in C_j} ||\boldsymbol{x} - \boldsymbol{c}_j||^2$$

where $\boldsymbol{c}_j$ is the mean of points in S_j.

The main steps of the algorithm are as follows.

i. A user indicates that data should be grouped into k clusters.

ii. Two initialization methods for the algorithm are commonly used: Forgy and Random Partition. The Forgy method randomly chooses k observations from the dataset and uses these as the initial means, $\boldsymbol{c}_j (1 \leq j \leq k)$. The Random Partition method first randomly assigns a cluster to each observation and then compute the initial mean, $\boldsymbol{c}_j (1 \leq j \leq k)$, to be the centroid of the cluster's randomly assigned points.

iii. *Assignment step*: Calculate the Euclidean distance between each data object $\boldsymbol{x}_i (1 \leq i \leq n)$ and all k cluster centers $\boldsymbol{c}_j (1 \leq j \leq k)$ and assign data object $\boldsymbol{x}_i$ to the nearest cluster (Fahim et al., 2006).

$$C_j^{(t)} = \{\boldsymbol{x}_i : ||\boldsymbol{x}_i - \boldsymbol{c}_j^{(t)}||^2 \leq ||\boldsymbol{x}_i - \boldsymbol{c}_{j'}^{(t)}||^2 \forall j, 1 \leq j \leq k$$

where each $\boldsymbol{x}_i$ is assigned to exactly one $C_j^{(t)}$.

If $\boldsymbol{x}_i$ is two-dimensional the Euclidean distance is given by

$$\mathrm{d}(\boldsymbol{x}_i, \boldsymbol{c}_j) = \left[\sum_{i=1}^{2}(\boldsymbol{x}_i - \boldsymbol{c}_j)^2\right]^{1/2}.$$

iv. *Update step*: For each cluster $j(1 \leq j \leq k)$, recalculate the new cluster center $\boldsymbol{c}_j$.

v. Repeat steps three and four until the cluster centers of all data do not change. According to Nazeer e Sebastian (2009) the K-means clustering algorithm always converge to local minimum.

The K-means algorithm is implemented in the `kmeans()` function from the `stats` package, in the software **R** (R Core Team, 2020). It allows the number of points in cluster is the Euclidean distance between point and cluster, this procedure is to search for a K-partition with locally optimal within-cluster sum of squares by moving points from one cluster to another (Hartigan e Wong, 1979).

Other methods that are under study to be compared with the K-means algorithm are Automatic Interaction Detection (AID) and random forests.

1.4 Analysis of the case-study - number of eggs

1.4.1 Poisson model

To analyse the number of eggs data obtained from the experiments described in Session 1.2, we first assume the simplest underlying process that the eggs are laid independently, singly, and at random at some constant underlying rate (Hinde e Demétrio, 1998; Demétrio et al., 2014). The response variable of interest, number of eggs, is simply a count, Y_{ij} and the Poisson distribution provides a starting point for data analysis.

We begin by fitting a standard Poisson log-linear model with the factors block and isolates as fixed effects, using the maximal linear predictor:

$$\eta_{ij} = \mu + \beta_j + \alpha_i, \quad j = 1, \ldots, 5, \quad i = 1, \ldots, 28, \tag{1.26}$$

where μ is the intercept, β_j is the fixed effect of j-th block, and α_i is the fixed effect of the i-th isolate. The analysis of deviance (Table 1.1) show that there is evidence from the residual deviance and X^2 values that the model does not fit to the data satisfactorily. This can also be seen in the half-normal plot of the deviance residuals with simulated envelope shown in (Figure 1.3 (A)). There is more variability ($\hat{\phi} > 1$) than the Poisson model accommodates, a clear evidence of overdispersion.

Models for overdispersed count data move away from the script Poisson assumption of equal mean and variance (dispersion index $\phi = 1$). As a next step we may try to

Table 1.1. Analysis of deviance (and X^2) for the number of eggs, using a Poisson log-linear model.

Experiment I					
Sources of variation	df	Deviance	p-value	X^2	p-value
Block	4	16.59			
Isolates	27	325.48			
Residual	108	400.45	<0.01	379.56	<0.01
$\hat{\phi} = 379.56/108 = 3.51$					
Experiment II - 60 days after root inoculation					
Sources of variation	df	Deviance	p-value	X^2	p-value
Block	4	22.29			
Isolates	27	415.77			
Residual	108	670.27	<0.01	620.47	<0.01
$\hat{\phi} = 620.47/108 = 5.75$					
Experiment II - 120 days after root inoculation.					
Sources of variation	df	Deviance	p-value	X^2	p-value
Block	4	24.07			
Isolates	27	186.55			
Residual	108	481.51	<0.01	449.08	<0.01
$\hat{\phi} = 449.08/108 = 4.16$					

accommodate the extra variability by considering approaches to allowing for overdispersion (Demétrio et al., 2014).

1.4.2 Quasi-Poisson model

The simplest way of taking overdispersion into account is to assume that the extra-dispersion is constant and independent of the number of eggs produced, replacing the variance function of the Poisson model by the more general form (1.3). Fitting a quasi-Poisson model with log link and the same linear predictor (1.26) to the number of eggs, Y_{ij}, the estimated values of $\phi = X^2/(\text{Res df})$ are given in Table 1.1.

The plots presented in (Figure 1.3(B)) show there is evidence of an adequate model fit, with most of the observed residuals lying within the simulated envelopes.

It is important to note the quasi-Poisson is based only on first and second moments assumptions and the drawback of this framework is that it does not provide an associated probability distribution. Next we present some alternative models that are distribution based.

1.4.3 Negative Binomial model

Assuming now that the eggs are laid not independently or at some varying underlying rate (differences in fertility of the *T. urticae* females), contributing additional

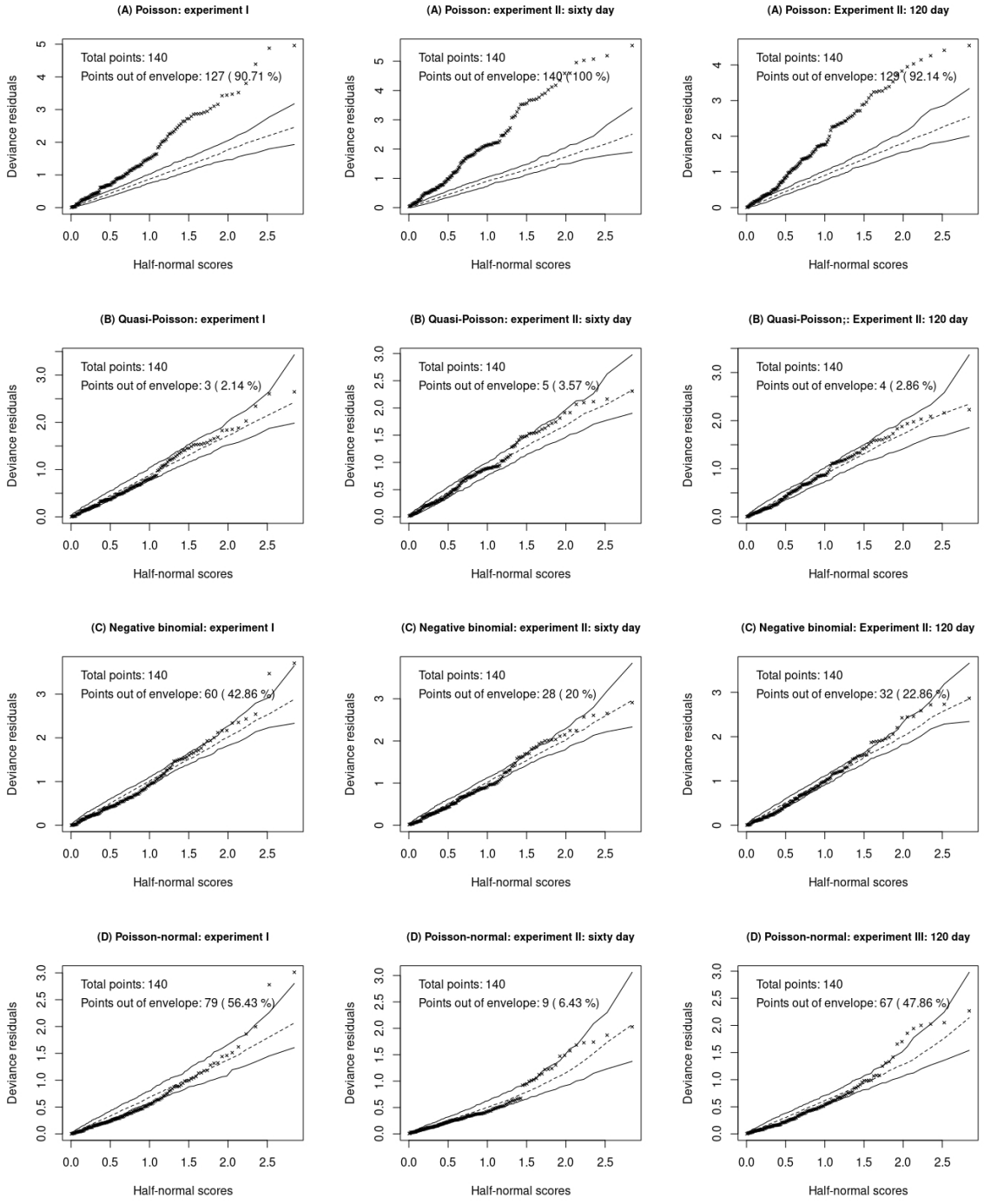

Figure 1.3. Half-normal plot with simulation envelopes for deviance residuals for (A) Poisson, (B) quasi-Poisson (C) Negative binomial and (D) Poisson-normal log-linear.

variability to the recorded counts, a two-stage model like the negative binomial model could give an explanation for the extra dispersion. Fitting the negative binomial model with log link and the same linear predictor (1.26) to the number of eggs, Y_{ij}, making use of the MASS package, the estimated values $\hat{\alpha}$ for all the experiments are $13.35(2, 64)$, $3.61(0.564)$, and $6.43(1.15)$ implying considerable overdispersion.

The half-normal plot presented in (Figure 1.3(C)) show that there is evidence for

all the experiments that the negative binomial model does not fit to the data, there is a considerable amount of points outside of the simulated envelopes.

1.4.4 Poisson-Normal model

Since we may think that there is a combination of many unexplained sources affecting the number of eggs, we can include a normal random effect at observation level, $Z_{ij} \sim N(0, \sigma_z^2)$, in the linear predictor,

$$\eta_{ij} = \mu + \beta_j + \alpha_i + Z_{ij}, \quad j = 1, \ldots, 5, \quad i = 1, \ldots, 28, \tag{1.27}$$

where Z_{ij} is a random effect with variance σ_z^2, β_j is the fixed effect of the j-th block and α_i is the fixed effect of the i-th isolate.

Fitting a Poisson-normal model, for these data the estimated values of σ_z^2 for all the experiments are $\sigma_{zI}^2 = 0,0729(0,2701)$, $\sigma_{zII60}^2 = 0,2856(0.5345)$ and $\sigma_{zII120}^2 = 0,1523(0.3903)$.

The half normal plot presented in (Figure 1.3(D)) show that there is evidence for all the experiments that the Poisson-normal model does not fit to the data, there is a considerable amount of points outside of the simulated envelopes.

Another way, instead of considering isolate as a fixed effect is to assume that it is a random effect, in the linear predictor,

$$\eta_{ij} = \mu + \beta_j + \alpha_i, \quad j = 1, \ldots, 5, \quad i = 1, \ldots, 28, \tag{1.28}$$

where α_i is a random effect with variance σ_I^2, that is $\alpha_i \sim N(0, \sigma_I^2)$.

Alternatively, we can assume both types of random effects, at observation level, $Z_{ij} \sim N(0, \sigma_z^2)$ and at isolate level, $\alpha_i \sim N(0, \sigma_I^2)$, in the linear predictor

$$\eta_{ij} = \mu + \beta_j + \alpha_i + Z_{ij}, \quad j = 1, \ldots, 5, \quad i = 1, \ldots, 28. \tag{1.29}$$

The half normal plots (not presented here) for the models with linear predictors given by equations (1.28) and (1.29) show that they also do not fit to the data.

1.4.5 Negative-binomial-normal model

Assuming now that overdispersion can be caused by varying mean and extra random variability simultaneously, a negative-binomial-normal can be fitted to the number of eggs, Y_{ij}. In a similar way as for the Poisson-Normal model we can include random effects at observation and/or isolate level, with the linear predictors given by (1.27), (1.28) and (1.29).

The half normal plot presented in (Figure 1.4) show that there is evidence for all the experiments that the Negative-binomial-normal model with a normal random effect at observation level does not fit to the data, there is a considerable amount of points outside of the simulated envelopes. Similar plots are obtained when considering a normal random effect at isolate level and at both levels.

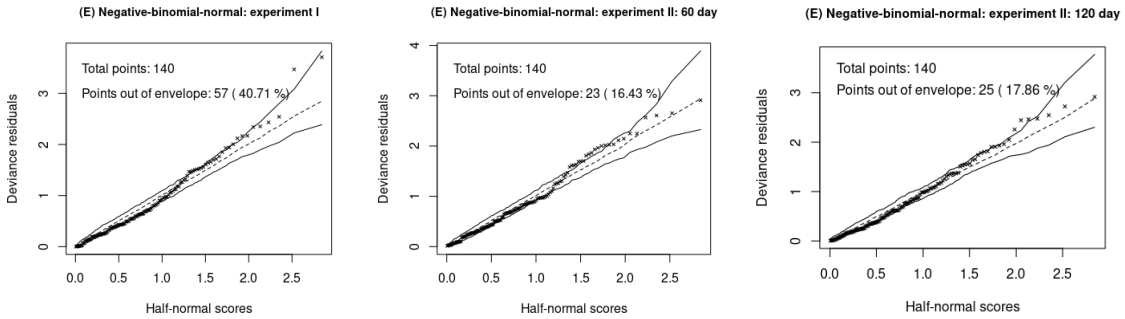

Figure 1.4. Half-normal plot with simulation envelopes for deviance residuals for (E) Negative-binomial-normal models

1.4.6 COM-Poisson model

Assuming now that for certain isolates the dispersion is smaller and for others is larger (see Figure 1.2B for experiment I) than predicted by the Poisson model, we can take into account underdispersion and overdispersion simultaneously, by fitting a COM-Poisson$_\mu$ model to the number of eggs, Y_{ij}, using the same linear predictor (1.26) for the mean and a constant dispersion or a regression for the dispersion with linear predictor

$$\eta_{ij} = \mu + \beta_j + \alpha_i, \qquad j = 1, \ldots, 5, \quad i = 1, \ldots, 28, \tag{1.30}$$

where μ is the intercept, β_J is the fixed effect of j-th block, and α_i is the fixed effect of the i-th isolate.

1.4.7 Estimates and model selection

The estimated values of the parameters and goodness-of-fit measures for the Poisson, Quasi-Poisson, Negative Binomial, Poisson-normal, Negative-Binomial-Normal (combined) and COM-Poisson$_\mu$ models for all the experiments are given in Tables 1.2, 1.3 and 1.4 while for the COM-Poisson$_\mu$ model with varying dispersion are in Table 1.5.

The results presented in Tables 1.2, 1.3 and 1.4 for the goodness-of-fit measures (log-likelihood, AIC and BIC) show that the COM-Poisson$_\mu$ model with constant dispersion gives the best fit for all experiments. The Poisson model is unsuitable, being conservative, due to the underestimated standard errors. The difference between the log-likelihood of the Poisson and COM-Poisson$_\mu$ model was 48.966 for one additional parameter, which confirms the significantly fit of the COM-Poisson$_\mu$ model. The estimated values ($\hat{\phi}$) of the constant dispersion parameter $-1, 133, -1.797, -1.299$, respectively for Experiment I, Experiment II with 60 days and Experiment II with 120 days confirms the overdispersion evidence.

The half-normal plots presented in Figures 1.3 and 1.4 show that among Poisson, Quasi-Poisson, Negative Binomial, Poisson-normal and Negative-Binomial-Normal mod-

Table 1.2. Parameter estimates (Est) and standard errors (SE) for the Poisson, Quasi-Poisson, Negative Binomial, Poisson-normal, Negative-Binomial-Normal (combined) and COM-Poisson$_\mu$ models. Experiment I.

Parameter	Poi	Q.-P.	Neg.-Bin	Poi-Nor	Neg.-Bin-Nor	CMP$_\mu$
ϕ		3.514	13.346		5017.82	-1,133
σ				0.073	0.073	
Intercept	3.689 (0.080)	3.689 (0.150)	3.705 (0.153)	3.677 (0.153)	3.681 (0.152)	3.718 (0.149)
1306	0.104 (0.102)	0.104 (0.191)	0.099 (0.201)	0.123 (0.199)	0.119 (0.199)	0.126 (0.196)
1451	-0.381 (0.116)	-0.381 (0.217)	-0.394 (0.208)	-0.374 (0.207)	-0.380 (0.207)	-0.400 (0.207)
1587	-0.405 (0.117)	-0.405 (0.219)	-0.410 (0.209)	-0.420 (0.208)	-0.425 (0.208)	-0.424 (0.208)
1604	-0.584 (0.123)	-0.584 (0.232)	-0.584 (0.212)	-0.575 (0.211)	-0.579 (0.211)	-0.603 (0.219)
1608	-0.172 (0.109)	-0.172 (0.205)	-0.181 (0.205)	-0.189 (0.204)	-0.194 (0.204)	-0.187 (0.198)
1610	-0.584 (0.123)	-0.584 (0.232)	-0.605 (0.213)	-0.606 (0.212)	-0.610 (0.212)	-0.605 (0.219)
1618	-0.373 (0.116)	-0.373 (0.217)	-0.384 (0.208)	-0.367 (0.207)	-0.372 (0.207)	-0.392 (0.207)
1622	-0.802 (0.133)	-0.803 (0.249)	-0.810 (0.218)	-0.841 (0.218)	-0.846 (0.219)	-0.823 (0.233)
1629	-0.068 (0.106)	-0.068 (0.199)	-0.068 (0.203)	-0.057 (0.202)	-0.062 (0.202)	-0.076 (0.195)
1634	-0.319 (0.114)	-0.319 (0.214)	-0.315 (0.207)	-0.309 (0.206)	-0.314 (0.206)	-0.336 (0.204)
1635	-0.397 (0.116)	-0.397 (0.218)	-0.427 (0.209)	-0.451 (0.209)	-0.456 (0.209)	-0.417 (0.208)
1636	-0.594 (0.124)	-0.594 (0.232)	-0.600 (0.213)	-0.658 (0.214)	-0.663 (0.214)	-0.614 (0.219)
1637	-0.296 (0.113)	-0.297 (0.212)	-0.301 (0.207)	-0.293 (0.205)	-0.298 (0.206)	-0.314 (0.203)
1638	-0.500 (0.120)	-0.500 (0.225)	-0.514 (0.211)	-0.554 (0.211)	-0.558 (0.211)	-0.519 (0.214)
1641	-0.755 (0.131)	-0.755 (0.245)	-0.763 (0.217)	-0.757 (0.216)	-0.761 (0.216)	-0.775 (0.230)
1669	-0.430 (0.118)	-0.430 (0.221)	-0.446 (0.209)	-0.432 (0.208)	-0.437 (0.208)	-0.450 (0.210)
1684	-0.491 (0.120)	-0.491 (0.225)	-0.497 (0.211)	-0.510 (0.210)	-0.514 (0.210)	-0.510 (0.213)
1709	-0.357 (0.115)	-0.357 (0.216)	-0.369 (0.208)	-0.350 (0.207)	-0.355 (0.207)	-0.376 (0.206)
3323	-0.365 (0.115)	-0.365 (0.216)	-0.368 (0.208)	-0.392 (0.208)	-0.397 (0.208)	-0.383 (0.206)
3375	-0.397 (0.116)	-0.397 (0.218)	-0.402 (0.209)	-0.382 (0.207)	-0.386 (0.207)	-0.416 (0.208)
3692	-0.575 (0.123)	-0.575 (0.231)	-0.581 (0.212)	-0.564 (0.211)	-0.568 (0.211)	-0.594 (0.218)
3693	-0.373 (0.116)	-0.373 (0.217)	-0.378 (0.208)	-0.364 (0.207)	-0.369 (0.207)	-0.391 (0.207)
3703	-0.166 (0.109)	-0.166 (0.205)	-0.181 (0.205)	-0.182 (0.204)	-0.186 (0.204)	-0.181 (0.198)
43	-0.699 (0.128)	-0.699 (0.240)	-0.715 (0.216)	-0.745 (0.216)	-0.750 (0.216)	-0.719 (0.226)
Nemix	0.166 (0.100)	0.166 (0.188)	0.155 (0.200)	0.171 (0.199)	0.167 (0.199)	0.212 (0.199)
Control	0.370 (0.096)	-0.370 (0.180)	0.368 (0.198)	0.382 (0.196)	-0.377 (0.197)	0.691 (0.268)
PL63	-0.005 (0.105)	-0.005 (0.196)	-0.013 (0.202)	0.011 (0.201)	-0.006 (0.201)	-0.007 (0.194)
Loglik	-552.4920	-	-505.3184	-506.6098	-506.6086	-493.1054
AIC	1168.984	-	1076.637	1079.220	1081.2017	1052.2109
BIC	1263.117	-	1173.711	1176.294	1181.233	1149.2851

els, the model that gave a best fit is the Quasi-Poisson. It was not possible to get the half-normal plot for the COM-Poisson$_\mu$ model due to convergence problems.

Comparing the COM-Poisson model with constant dispersion (Tables 1.2, 1.3 and 1.4) and the COM-Poisson model with varying dispersion (Table 1.5), using the likelihood ratio test we can see that of Com-Poisson model with varying dispersion gives a better fit (2*(−433.4589 + 493.1054) = 59.6465 = 119.293 for Experiment I; 2*(−471.4465 + 486.418) = 2*14.9715 = 29.943 for Experiment II with 60 days and 2*(−457.0709 + 477.976) = 2*20.9051 = 41.8102 for Experiment II with 120 days).

1.4.8 Grouping means

The predicted means and respective standard errors for the number of eggs for the well-fitted Quasi-Poisson model were used because of its simplicity, to obtain the groups of the isolates.

Table 1.3. Parameter estimates (Est) and standard errors (SE) for Poisson, Quasi-Poisson, Negative Binomial, Poisson-normal, Negative-Binomial-Normal (combined) and COM-Poisson$_\mu$ models. Experiment II with 60 days.

Parameter	Poi	Q.-P.	Neg.-Bin	Poi-Nor	Neg.-Bin-Nor	CMP$_\mu$
ϕ		5.745	3.614		2916.049	-1.797
σ			0.564	0.286	0.285	
Intercept	2.539 (0.132)	2.539 (0.316)	2.541 (0.284)	2.425 (0.292)	2.426 (0.293)	2.517
1306	0.754 (0.151)	0.754 (0.363)	0.762 (0.365)	0.822 (0.374)	0.821 (0.375)	0.785
1451	0.514 (0.158)	0.514 (0.379)	0.522 (0.368)	0.588 (0.377)	0.588 (0.377)	0.530
1587	0.248 (0.167)	0.248 (0.400)	0.248 (0.372)	0.252 (0.382)	0.251 (0.383)	0.255
1604	0.560 (0.157)	0.560 (0.375)	0.564 (0.368)	0.570 (0.377)	0.569 (0.378)	0.577
1608	0.104 (0.172)	0.104 (0.413)	0.075 (0.375)	-0.038 (0.390)	-0.039 (0.391)	0.098
1610	0.284 (0.165)	0.284 (0.397)	0.260 (0.372)	0.241 (0.383)	0.241 (0.384)	0.286
1618	-0.901 (0.232)	-0.901 (0.557)	-0.918 (0.407)	-0.891 (0.415)	0.892 (0.418)	-0.908
1622	0.172 (0.170)	0.172 (0.407)	0.191 (0.373)	0.177 (0.383)	0.176 (0.385)	-0.182
1629	0.260 (0.166)	0.260 (0.399)	0.214 (0.373)	0.064 (0.388)	0.063 (0.390)	0.257
1634	0.523 (0.158)	0.523 (0.378)	0.522 (0.368)	0.552 (0.377)	0.552 (0.378)	0.538
1635	0.352 (0.163)	0.352 (0.391)	0.340 (0.371)	0.311 (0.382)	0.310 (0.383)	0.360
1636	0.374 (0.162)	0.374 (0.389)	0.393 (0.370)	0.389 (0.380)	0.389 (0.381)	0.388
1637	0.223 (0.168)	0.223 (0.402)	0.206 (0.373)	0.170 (0.384)	0.169 (0.385)	0.225
1638	-1.023 (0.243)	-1.023 (0.583)	-1.034 (0.413)	-1.037 (0.423)	-1.038 (0.425)	-1.025
1641	-0.267 (0.190)	-0.267 (0.455)	-0.260 (0.383)	-0.256 (0.393)	-0.257 (0.394)	-0.270
1669	-0.288 (0.191)	-0.288 (0.458)	-0.308 (0.384)	-0.327 (0.396)	-0.327 (0.397)	-0.305
1684	0.272 (0.166)	0.272 (0.398)	0.287 (0.372)	0.077 (0.388)	0.077 (0.389)	0.282
1709	0.248 (0.167)	0.248 (0.400)	0.235 (0.372)	0.181 (0.384)	0.180 (0.385)	0.251
3323	-0.330 (0.193)	-0.330 (0.463)	-0.362 (0.386)	-0.443 (0.400)	-0.444 (0.402)	-0.358
3375	0.272 (0.166)	0.272 (0.398)	0.295 (0.371)	-0.140 (0.386)	0.140 (0.387)	0.284
3692	0.185 (0.169)	0.185 (0.405)	0.175 (0.373)	0.245 (0.382)	0.244 (0.383)	0.188
3693	-0.151 (0.184)	-0.151 (0.441)	-0.187 (0.381)	-0.150 (0.390)	-0.150 (0.392)	-0.170
3703	0.157 (0.170)	0.159 (0.408)	0.133 (0.374)	0.143 (0.384)	0.143 (0.385)	0.157
43	0.363 (0.163)	0.363 (0.390)	0.327 (0.371)	0.280 (0.383)	0.280 (0.384)	0.366
Nemix	0.965 (0.147)	0.965 (0.352)	0.943 (0.364)	1.028 (0.372)	1.027 (0.373)	1.031
Control	1.093 (0.144)	1.093 (0.346)	1.102 (0.363)	1.184 (0.371)	1.184 (0.371)	1.212
PL63	-0.032 (0.178)	-0.032 (0.427)	-0.019 (0.377)	0.030 (0.386)	0.029 (0.387)	-0.029
Loglik	-640.310	-	-492.868	-497.836	-497.832	-486.418
AIC	1344.6	-	1051.7	1061.7	1063.7	1038.837
BIC	1263.117	-	1173.711	1158.7	1163.7	1135.911

Empirical grouping - Visual inspection

As suggested by Fatoretto et al. (2018), the isolates can be grouped by visual inspection using a likelihood-ratio test to identify the number of groups based on the similarity observed between isolate predicted means (Figure 1.5).

For the analysis of number of eggs, using the isolate predicted means, by visual inspection, we started creating ten groups of similar isolates for all experiment as shown in Table 1.6 and Figure 1.5. Table 1.8 shows the values for residual deviances for models with different numbers of groups and we used the F-test to get up to four groups of isolates.

We first compared the model with 28 isolates with the model with ten groups for all experiment Table 1.6 and Figure 1.5, resulting in a non significant F-test. Similarities between the isolate groups were searched by merging groups, with the aim of reducing the number of groups. The values of the residual deviances for the different groupings are presented in Table 1.8.

For experiment I, a first reduced Grouping 2 was created by merging Control and

Table 1.4. Parameter estimates (Est) and standard errors (SE) for the Poisson, Quasi-Poisson, Negative Binomial, Poisson-normal, Negative-Binomial-Normal (combined) and COM-Poisson$_\mu$ models. Experiment II with 120 days.

Parameter	Poi	Q.-P.	Neg.-Bin	Poi-Nor	Neg.-Bin-Nor	CMP$_\mu$
ϕ		4.157	6.433		3179.376	-1.299
σ			1.15	0.152	0.152	0.137
Intercept	2.984 (0.115)	2.984 (0.234)	2.982 (0.221)	2.887 (0.224)	2.888 (0.225)	2.984 (0.212)
1306	0.511 (0.138)	0.511 (0.281)	0.528 (0.285)	0.562 (0.287)	0.561 (0.288)	0.517 (0.257)
1451	0.154 (0.149)	0.154 (0.303)	0.149 (0.291)	0.233 (0.292)	0.233 (0.292)	0.154 (0.275)
1587	0.112 (0.150)	0.112 (0.306)	0.122 (0.291)	0.176 (0.293)	0.175 (0.293)	0.113 (0.277)
1604	0.023 (0.153)	0.023 (0.313)	0.017 (0.293)	0.068 (0.295)	0.066 (0.296)	0.022 (0.283)
1608	0.174 (0.148)	0.174 (0.302)	0.179 (0.290)	0.263 (0.291)	0.261 (0.292)	0.175 (0.274)
1610	-0.036 (0.156)	-0.036 (0.317)	-0.016 (0.294)	0.051 (0.295)	0.050 (0.296)	-0.035 (0.287)
1618	-0.140 (0.160)	-0.140 (0.326)	-0.158 (0.297)	-0.103 (0.299)	-0.104 (0.300)	-0.145 (0.295)
1622	-0.581 (0.182)	-0.581 (0.371)	-0.613 (0.310)	-0.603 (0.314)	-0.605 (0.315)	-0.599 (0.332)
1629	-0.024 (0.155)	-0.024 (0.316)	-0.030 (0.294)	0.001 (0.297)	0.001 (0.297)	-0.025 (0.286)
1634	-0.304 (0.167)	-0.304 (0.341)	-0.291 (0.300)	-0.291 (0.304)	-0.292 (0.305)	-0.305 (0.306)
1635	-0.241 (0.164)	-0.241 (0.335)	-0.234 (0.299)	-0.211 (0.302)	-0.212 (0.302)	-0.244 (0.302)
1636	-0.370 (0.171)	-0.370 (0.348)	-0.380 (0.303)	-0.388 (0.307)	-0.390 (0.308)	-0.377 (0.312)
1637	-0.036 (0.156)	-0.036 (0.317)	-0.063 (0.295)	-0.049 (0.298)	-0.049 (0.299)	-0.040 (0.287)
1638	-0.127 (0.159)	-0.127 (0.325)	-0.137 (0.296)	-0.066 (0.298)	-0.067 (0.299)	-0.130 (0.293)
1641	0.035 (0.153)	0.035 (0.312)	0.024 (0.293)	0.042 (0.296)	0.040 (0.297)	0.033 (0.282)
1669	-0.304 (0.167)	-0.304 (0.341)	-0.298 (0.300)	-0.302 (0.304)	-0.303 (0.305)	0.306 (0.306)
1684	0.012 (0.154)	0.012 (0.314)	0.017 (0.293)	0.085 (0.295)	0.084 (0.295)	0.011 (0.284)
1709	0.144 (0.149)	0.144 (0.304)	0.158 (0.290)	0.238 (0.292)	0.237 (0.292)	0.146 (0.275)
3323	0.213 (0.147)	0.213 (0.299)	0.199 (0.290)	0.272 (0.291)	0.271 (0.292)	0.213 (0.272)
3375	-0.140 (0.160)	-0.140 (0.326)	-0.145 (0.297)	-0.083 (0.298)	-0.084 (0.299)	-0.143 (0.294)
3692	-0.087 (0.158)	-0.087 (0.322)	-0.060 (0.295)	-0.049 (0.298)	-0.050 (0.298)	-0.085 (0.290)
3693	0.080 (0.151)	0.080 (0.308)	0.089 (0.292)	0.110 (0.294)	0.108 (0.295)	0.080 (0.279)
3703	0.444 (0.140)	0.444 (0.285)	0.451 (0.286)	0.547 (0.286)	0.546 (0.287)	0.449 (0.260)
43	0.233 (0.146)	0.233 (0.298)	0.223 (0.289)	0.236 (0.292)	0.235 (0.293)	0.233 (0.270)
Nemix	0.546 (0.137)	0.546 (0.279)	0.538 (0.285)	0.592 (0.287)	0.591 (0.287)	0.552 (0.256)
Control	0.553 (0.137)	0.553 (0.279)	0.553 (0.285)	0.648 (0.286)	0.648 (0.286)	0.560 (0.256)
PL63	-0.074 (0.157)	-0.074 (0.321)	-0.050 (0.294)	0.004 (0.296)	0.002 (0.297)	-0.072 (0.289)
Loglik	-559.417	-	-485.746	-488.477	-488.474	-477.976
AIC	1182.8	-	1037.5	1043.0	1044.9	1021.953
BIC	1276.967	-	1134.565	1140.0	1145.0	1119.027

isolates from G II, giving a non significant result. Grouping 3 was created by merging Control and isolates from G II and G III, giving a non significant test. Grouping 4 was created by, additionally to Grouping 2, merging isolates from G III and G IV giving a non-significant test. Grouping 5 was created by, additionally, to Grouping 2, merging isolates from G III, G IV and G V giving a non significant test. Grouping 6 was created by, additionally to Grouping 4, merging isolates from G V and G VI, giving a non-significant test. Grouping 7 was created additionally to Grouping 4, merging isolates from G V, G VI and G VII, giving a non significant test. Finally, Grouping 8 was created additionally to Grouping 4, merging isolates from G VII and G VIII, giving a significant test. This process resulted in four groups of isolates as shown in Table 1.7 and Figure 1.6. A half-normal plot for the deviance residuals after fitting a quasi-Poisson with four groups of isolates confirms the evidence of a well- fitted model (Figure 1.7). The four groups (Table 1.7) would be classified as

- group I: highly promising isolates;
- group II: moderately promising isolates;

Table 1.5. Parameter estimates (Est) and standard error (SE) for the fitted COM-Poisson model with dispersion for all experiments.

	Exp. I		Exp. II 60 days		Exp. II 120 days	
Parameter	Mean	Dispersion	Mean	Dispersion	Mean	Dispersion
Intercept	3.600 (0.118)	-0.921 (0.701)	2.411 (0.223)	-1.000 (0.738)	3.061 (0.175)	-1.000 (0.401)
1306	0.100 (0.118)	9.434 (1.971)	0.768 (0.233)	0.333 (1.039)	0.511 (0.256)	-0.682 (0.800)
1451	-0.381 (0.147)	0.911 (0.979)	0.528 (0.232)	0.633 (1.045)	0.145 (0.186)	1.914 (0.678)
1587	-0.398 (0.225)	-0.676 (0.964)	0.259 (0.269)	-0.013 (1.010)	0.093 (0.309)	-1.000 (1.027)
1604	-0.584 (0.145)	1.203 (0.924)	0.576 (0.311)	-1.000 (1.081)	0.023 (0.228)	0.342 (0.736)
1608	-0.173 (0.188)	-0.330 (1.044)	0.110 (0.348)	-1.000 (0.990)	0.165 (0.195)	1.319 (0.719)
1610	-0.580 (0.204)	-0.179 (1.013)	0.284 (0.313)	-0.688 (1.112)	-0.037 (0.243)	0.110 (0.871)
1618	-0.378 (0.141)	1.134 (1.135)	-0.885 (0.269)	1.199 (0.988)	-0.129 (0.224)	0.610 (0.742)
1622	-0.818 (0.312)	-1.277 (1.180)	0.182 (0.325)	-0.769 (1.123)	-0.639 (0.379)	-1.000 (1.642)
1629	-0.066 (0.141)	0.816 (0.913)	0.243 (0.343)	-1.000 (0.955)	-0.042 (0.240)	0.185 (0.732)
1634	-0.315 (0.157)	0.523 (0.915)	0.538 (0.301)	-0.818 (1.113)	-0.298 (0.304)	-0.475 (0.873)
1635	-0.413 (0.267)	-1.154 (1.090)	0.357 (0.330)	-1.000 (1.099)	-0.171 (0.309)	-0.662 (0.824)
1636	-0.640 (0.337)	-2.071 (——-)	0.388 (0.318)	-0.891 (1.132)	-0.342 (0.347)	-1.000 (0.910)
1637	-0.295 (0.162)	0.365 (0.915)	0.220 (0.342)	-1.000 (1.128)	-0.006 (0.295)	-0.671 (0.842)
1638	-0.512 (0.317)	-1.670 (1.215)	-1.000 (0.356)	0.166 (1.195)	-0.142 (0.200)	1.385 (0.720)
1641	-0.755 (0.196)	0.124 (0.947)	-0.257 (0.308)	-0.073 (1.066)	0.017 (0.235)	0.206 (0.674)
1669	-0.431 (0.172)	0.242 (0.961)	-0.295 (0.345)	-0.460 (1.199)	-0.273 (0.341)	-1.000 (0.920)
1684	-0.513 (0.238)	-0.713 (0.989)	0.264 (0.339)	-1.000 (0.910)	0.002 (0.277)	-0.465 (0.956)
1709	-0.356 (0.138)	1.243 (1.052)	0.240 (0.342)	-1.000 (1.124)	0.153 (0.215)	0.568 (0.839)
3323	-0.363 (0.241)	-0.915 (1.019)	-0.322 (0.374)	-1.000 (1.085)	0.196 (0.234)	0.054 (0.812)
3375	-0.395 (0.123)	2.575 (0.964)	0.291 (0.333)	-1.000 (0.906)	-0.140 (0.284)	-0.420 (0.931)
3692	-0.573 (0.121)	2.995 (1.461)	0.200 (0.214)	1.940 (1.075)	-0.102 (0.325)	-1.000 (0.883)
3693	-0.373 (0.147)	0.890 (0.905)	-0.140 (0.286)	0.121 (1.079)	0.038 (0.313)	-1.000 (0.816)
3703	-0.162 (0.188)	-0.338 (1.055)	0.169 (0.271)	0.048 (1.020)	0.412 (0.238)	-0.223 (1.162)
43	-0.703 (0.315)	-1.440 (1.207)	0.367 (0.329)	-1.000 (1.031)	0.186 (0.303)	-1.000 (0.764)
Nemix	0.168 (0.147)	0.343 (1.052)	0.979 (0.224)	0.504 (1.111)	0.505 (0.226)	-0.068 (0.738)
Control	0.371 (0.154)	-0.084 (0.919)	1.108 (0.224)	0.348 (1.061)	0.547 (0.175)	5.732 (1.372)
PL63	-0.003 (0.127)	1.612 (0.922)	-0.017 (0.231)	1.237 (1.011)	-0.082 (0.260)	-0.126 (0.882)
Loglik	−433.4589		−471.4465		−457.0709	
AIC	1026.918		1062.893		1034.142	
BIC	1203.416		1239.392		1210.640	

- group III: less promising isolates;
- group IV: isolates with no effect (similar to the worst Control).

For experiment II with 60 days, a first reduced Grouping 2 was created by merging Control and isolates from G II, giving a non significant result. Grouping 3 was created by, additionally, merging control, isolates of G II and G III, giving a non significant result. Grouping 4 was created by, additionally to Grouping 2, merging isolates from G III and G IV giving a non-significant test. Grouping 5 was created by, additionally, to Grouping 2, merging isolates from G III, G IV and G V giving a non-significant test. Grouping 6 was created by, additionally to Grouping 4, merging isolates from G III, G IV, G V and G VI, giving a non-significant test. Grouping 7 was created by, additionally to Grouping 4, merging isolates from G III, G IV, G V, G VI and G VII, giving a non significant test. Finally, grouping 8 was created additionally to Grouping 4, merging isolates from G VII and G VIII, giving a significant test. This process resulted in four groups of isolates as shown in Table 1.7 and Figure 1.6. A half-normal plot for the deviance residuals after fitting a quasi-Poisson with four groups of isolates confirms the evidence of a well- fitted model (Figure 1.7). The four groups (Table 1.7) would be classified as

- group I: highly promising isolates;

Table 1.6. First grouping according to the empirical grouping of the isolate predicted means.

Groups	Experiment I				
G I	1622				
G II	1641				
G III	43				
G IV	1636	1610	1604	3692	
G V	1638	1684			
G VI	1669	1587	1635	3375	1451
	3693	1618	3323	1709	1634
G VII	1608	3703			
G VIII	1629	PL63	1296		
G IX	1306	Quar			
G X	Cont				
Groups	Experiment II - 60				
G I	1638	1618			
G II	3323	1669	1641		
G III	3693				
G IV	PL23	1296	1608		
G V	3703	1622	3692	1637	1587
	1709	1629	3375	1684	1610
	1610				
G VI	1635	43	1636		
G VII	1451	1634	1604		
G VIII	1306				
G IX	Quar				
G X	Cont				
Groups	Experiment II - 120				
G I	1622				
G II	1636				
G III	1669	1634	1635		
G IV	1618	3375	1638	3692	PL23
G V	1610	1637	1629	1296	1684
	1604	1641			
G VI	3693	1587	1709	1451	
G VII	1608	3323	43		
G VIII	3703				
G IX	1306				
G X	Quar	Cont			

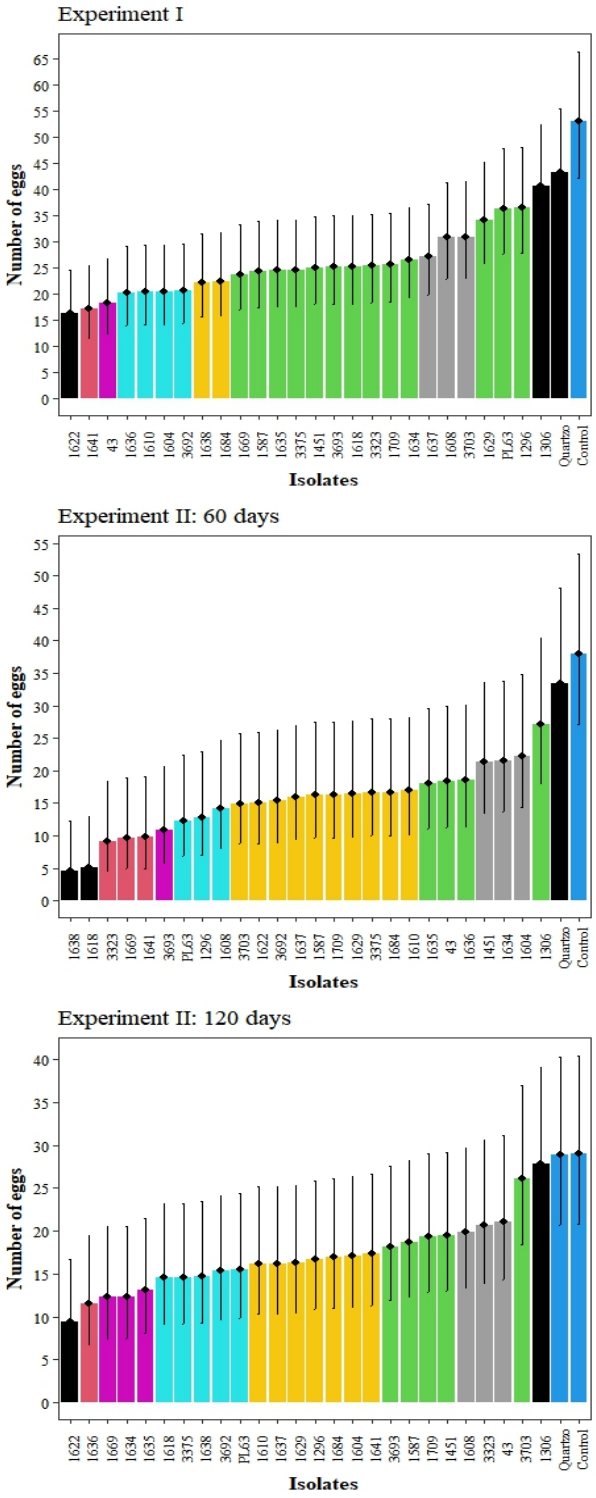

Figure 1.5. Plot with the predicted values

- group II: moderately promising isolates;
- group III: less promising isolates;
- group IV: isolates with no effect (similar to the worst Control).

For experiment II with 120 days, a first reduced Grouping 2 was created by

merging G I and GII, giving a non significant result. Grouping 3 was created by merging G I, GII and G III, giving a non-significant result. Grouping 4 was created by merging G I, GII, G III and G IV, giving a non significant result. Grouping 5 was created by, additionally to Grouping 3, merging isolates from G IV and G V giving a non-significant test. Grouping 6 was created by, additionally to Grouping 3, merging isolates from G IV, G V and G VI giving a non-significant test. Grouping 7 was created by, additionally to Grouping 3, merging isolates from G IV, G V, G VI and G VII giving a non significant test. Finally, additionally to Grouping 8, merging isolates from G VII and G VIII giving a significant test. This process resulted in four groups of isolates as shown in Table 1.7 and Figure 1.6. A half-normal plot for the deviance residuals after fitting a quasi-Poisson with four groups of isolates confirms the evidence of a well- fitted model (Figure 1.7). The four groups (Table 1.7) would be classified as

- group I: highly promising isolates;
- group II: moderately promising isolates;
- group III: less promising isolates;
- group IV: isolates with no effect (similar to the worst Control).

Table 1.8. Values of residual deviances for models with different numbers of groups

Number of		Empirical				K-means		
Groups	Df	Exp I	E II - 60	E II - 120	Df	E I	E II - 60	E II - 120
28 groups	108	400.75	670.27	481.51	108	400.75	670.27	481.51
10 groups	126	404.29	672.84	482.90	126	402.20	671.93	482.59
9 groups	127	404.38	673.63	483.24	127	402.70	673.63	482.96
8 groups	128	404.76	678.51	483.54	128	403.83	675.11	483.39
7 groups	129	405.89	682.83	484.26	129	405.15	678.16	484.43
6 groups	130	410.37	692.35	487.48	130	409.34	684.05	485.30
5 groups	131	419.16	693.82	488.68	131	415.69	688.61	488.46
4 groups	132	428.45	710.70	491.47	132	415.69	715.24	491.47
3 groups	133	458.19	756.71	508.57	133	449.20	742.80	506.74

K-means clustering

For the analysis of number of eggs, using the predicted means and respective standard errors, we started with $K = 10$ groups (Table 1.9 and Figure 1.8) and tested up to get four groups of isolates. For experiment I and II with 120 days we obtained four groups (Table 1.10 and Figure 1.9)

- group I: highly promising isolates;

Table 1.7. Final grouping according to the empirical grouping of the isolate predicted means.

Groups	Experiment I				
G I	1622	1641	43	1636	1610
	1604	3692	1638	1684	
G II	1669	1587	1635	3375	1451
	3693	1618	3323	1709	1634
	1637				
G III	1608	3703	1629	PL63	1296
G IV	Quar	1306	Cont		
Groups	Experiment II - 60				
G I	1638	1618			
G II	3323	1669	1641	3693	PL63
	1296				
G III	1608	3703	1622	3692	1637
	1587	1709	1629	3375	1684
	1610	1635	43	1636	1451
	1634	1604			
G IV	1306	Quar	Cont		
Groups	Experiment II - 120				
G I	1622	1636	1669	1634	1635
G II	1618	3375	1638	3692	PL63
	1610	1637	1629	1296	1684
	1604	1641			
G III	3693	1587	1709	1451	1608
	3323	43			
G IV	3703	1306	Quar	Cont	

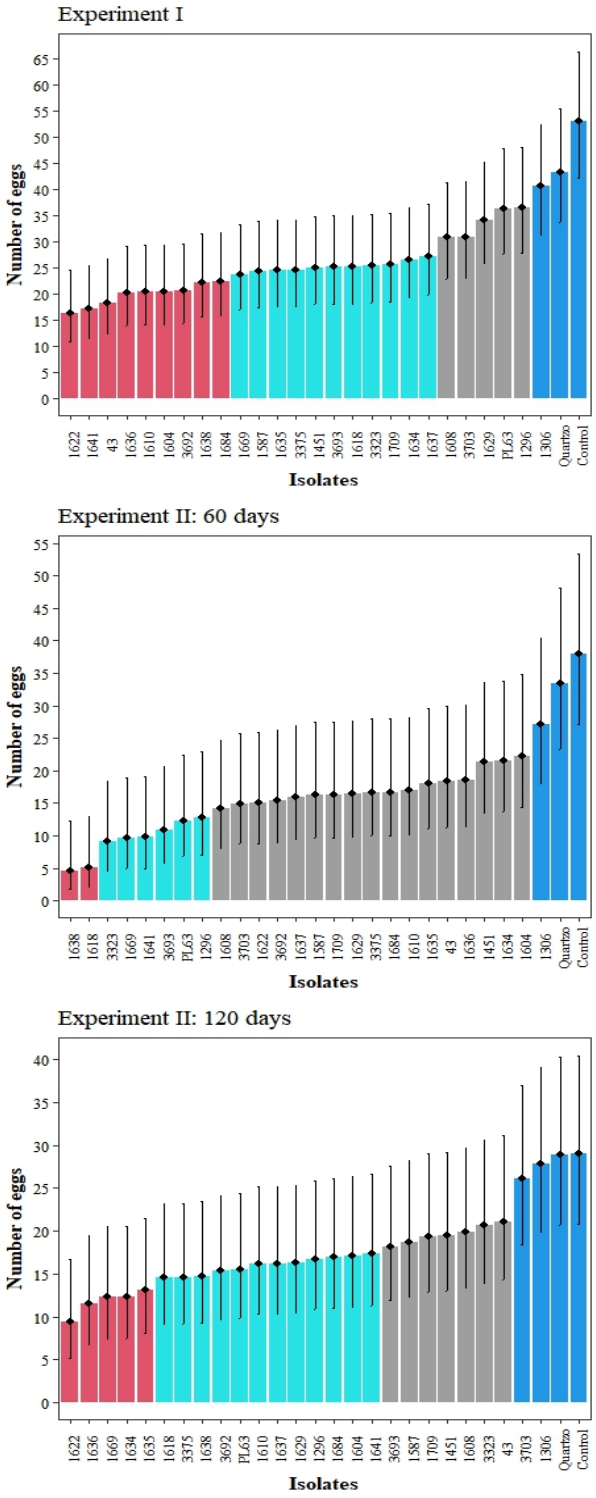

Figure 1.6. Plot with the predicted values

- group II: moderately promising isolates;
- group III: less promising isolates;
- group IV: isolates with no effect (similar to the worst control)

while for experiment II with 60 days we obtained five groups (Table 1.10 and Figure 1.9).

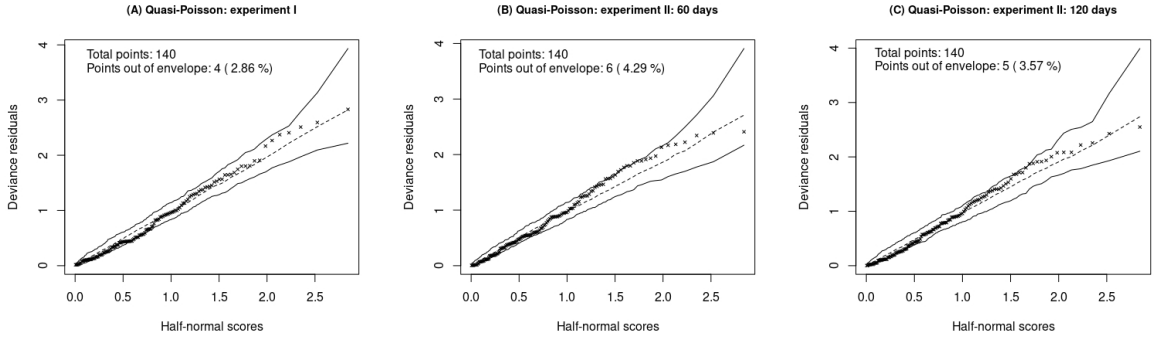

Figure 1.7. Half-normal plot with simulation envelopes of deviance residuals component for fitted groups using Quasi-Poisson model. Empirical grouping.

Half-normal plot for the deviance residuals after fitting a quasi-Poisson with four groups of isolates confirms the evidence of a well- fitted model (Figure 1.10). The four groups (Table 1.10) would be classified as

- group I: highly promising isolates;
- group II: less highly promising isolates;
- group III: moderately promising isolates;
- group IV: less promising isolates;
- group V: isolates with no effect (similar to the worst control).

1.4.9 Discussion

In this Section, we proposed different models that take into account overdispersion (underdispersion) to analyse the number of *T. urticae* eggs at 60 days and 120 days after root inoculation of strawberry plants inoculated with different promising isolates of the entomopathogenic fungi of *Metarhizium spp.*, *B. bassiana*, *I. fumosorosea*. We compared the results and also discussed model selection and diagnostics. For grouping the isolates we proposed two different methods. All the methods were implemented in the software R (R Core Team, 2020) and the scripts developed are presented in the Appendix.

Table 1.9. First grouping according to the nearby between the predicted values and standard error using the kmeans method.

Groups		Experiment I			
G I	1622	1641	43		
G II	1636	1610	1604	3692	
G III	1638	1684			
G IV	1669	1587	1635	3375	1451
	3693	1618	3323	1709	
G V	1634	1637			
G VI	1608	3703			
G VII	1629				
G VIII	PL63	1296			
G IX	1306	Quar			
G X	Cont				
Groups		Experiment II - 60			
G I	1638	1618			
G II	3323	1669	1641		
G III	3693	PL63	1296		
G IV	1608	3703	1622	3692	
G V	1637	1587	1709	1629	3375
	1684	1610			
G VI	1635	43	1636		
G VII	1451	1634	1604		
G VIII	1306				
G IX	Quar				
G X	Cont				
Groups		Experiment II - 120			
G I	1622				
G II	1636	1669	1634	1635	
G III	1618	3375	1638		
G IV	3692	PL23	1610	1637	1629
G V	1296	1684	1604	1641	
G VI	3693	1587			
G VII	1709	1451	1608		
G III	3323	43			
G IX	3703				
G X	1306	Quar	Cont		

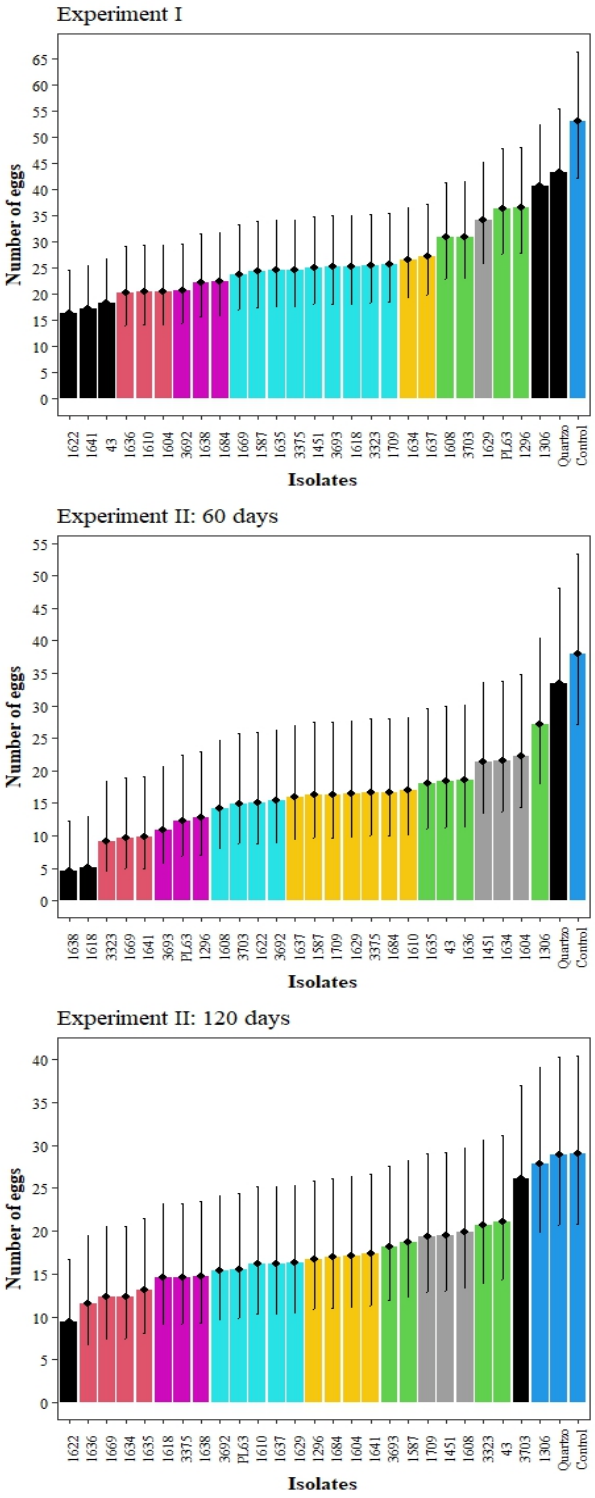

Figure 1.8. Plot with the predicted values

1.5 Analysis the case-study - number of flowers

The motivating dataset of this work had as one of its aims to evaluate the effect of promising entomopathogenic fungi inoculated on the roots of strawberry plants to control the population of mitesand, also, to evaluate how the fungi affect the development of the

Table 1.10. Final grouping chosen according to the proximity between the predicted values and the standard error. Kmeans method.

Groups			Experiment I		
G I	1622	1641	43	1636	1604
	1610	3692			
G II	1638	1684	1669	1587	1635
	3375	1451	1618	3693	3323
	1709	1634	1637		
G III	1608	3703	1629	PL63	1296
G IV	1306	Quar	Cont		
Groups			Experiment II - 60		
G I	1638	1618			
G II	3323	1669	1641	3693	PL63
	1296				
G III	1608	3703	1622	3692	1637
	1587	1709	1629	1684	3375
	1610	1635	43	1636	
G IV	1451	1634	1604	1306	
G V	Quar	Cont			
Groups			Experiment II - 120		
G I	1622	1636	1634	1669	1635
G II	1618	3375	1638	3692	PL63
	1610	1637	1629	1296	1684
	1604	1641			
G III	3693	1587	1709	1451	1608
	3323	43			
G IV	3703	1306	Quar	Cont	

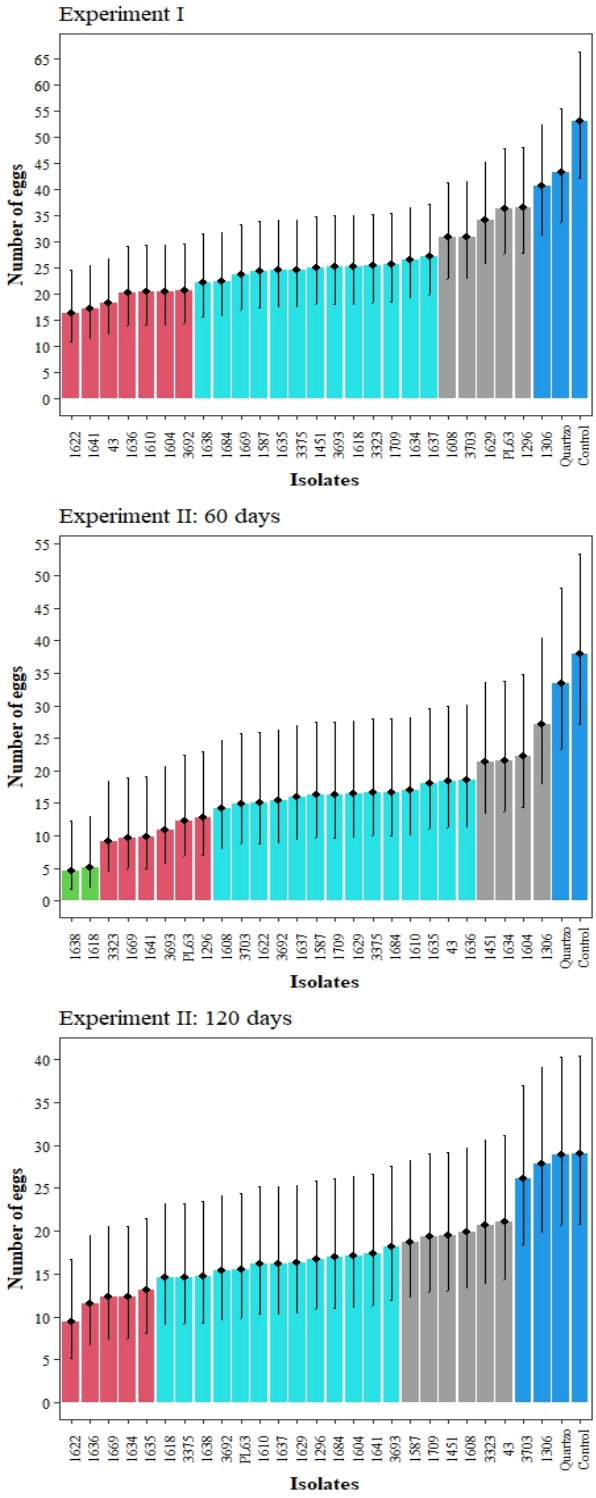

Figure 1.9. Plot with the predicted values

plants, mainly number of flowers and leaves, as a measure for the plants growth.

1.5.1 Exploratory analysis

The dispersion plots of the number of flowers for each isolates, over time for experiments I and II (Figures 1.11) show an increasing trend over time and that there

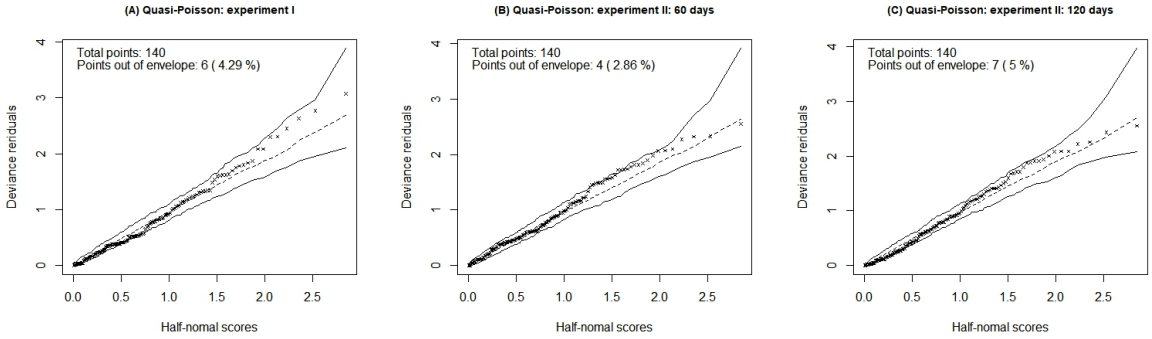

Figure 1.10. Half-normal plot with simulation envelopes of deviance residuals component for fitted groups using Quasi-Poisson model. Kmeans method.

are clear differences in the influence of the isolates and evidence of differing degrees of variability between replicates.

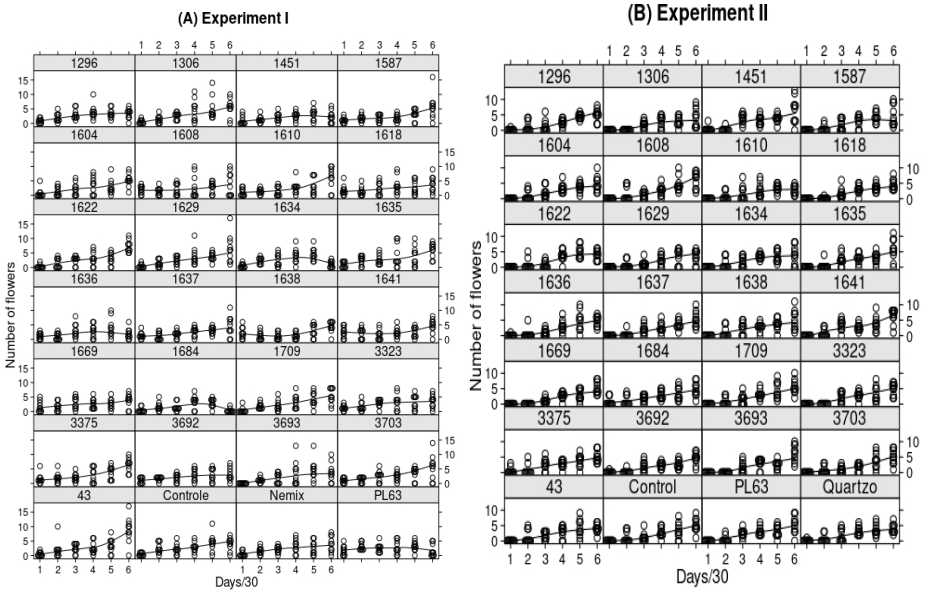

Figure 1.11. Dispersion plots of number of flowers per day versus days for (A) experiment I and (B) experiment II.

The dispersion plots of the sample variance versus sample means (Figures 1.12) show that there points below and above the identity line, suggesting evidence of overdispersion and underdispersion. Bar plots of the observed numbers of flowers for experiment I (Figure 1.13(A)) and experiment II (Figure 1.13(B)), suggests that there is zero-inflation, mainly for experiment II.

1.5.2 Poisson model

We begin by fitting a Poisson log-linear model with the factors block, isolate, and day as fixed effects, using the maximal linear predictor given by the equation (1.31).

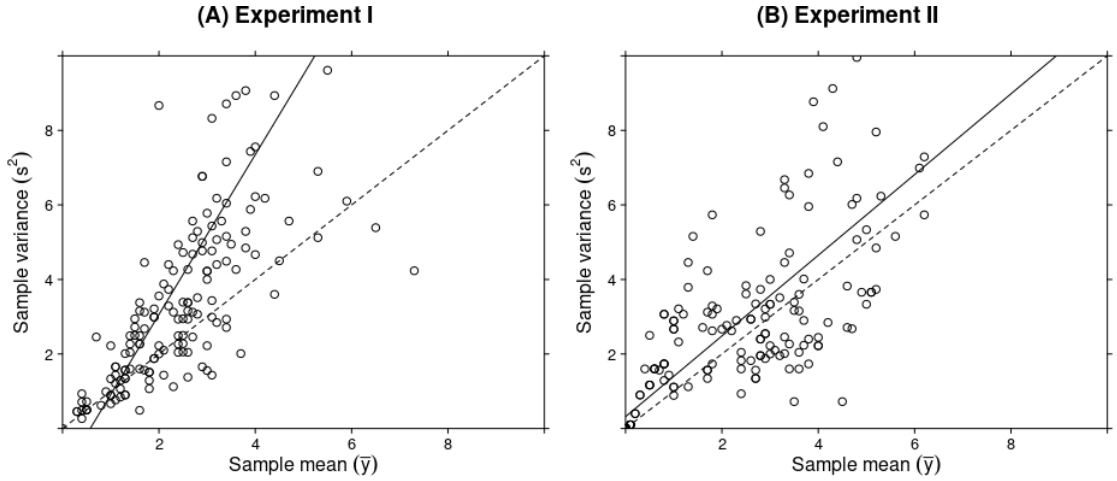

Figure 1.12. Dispersion plots sample mean versus the sample variance of number of flowers per day for (A) experiment I and (B) experiment II (dotted line is the identity line and the solid line is the least squares line).

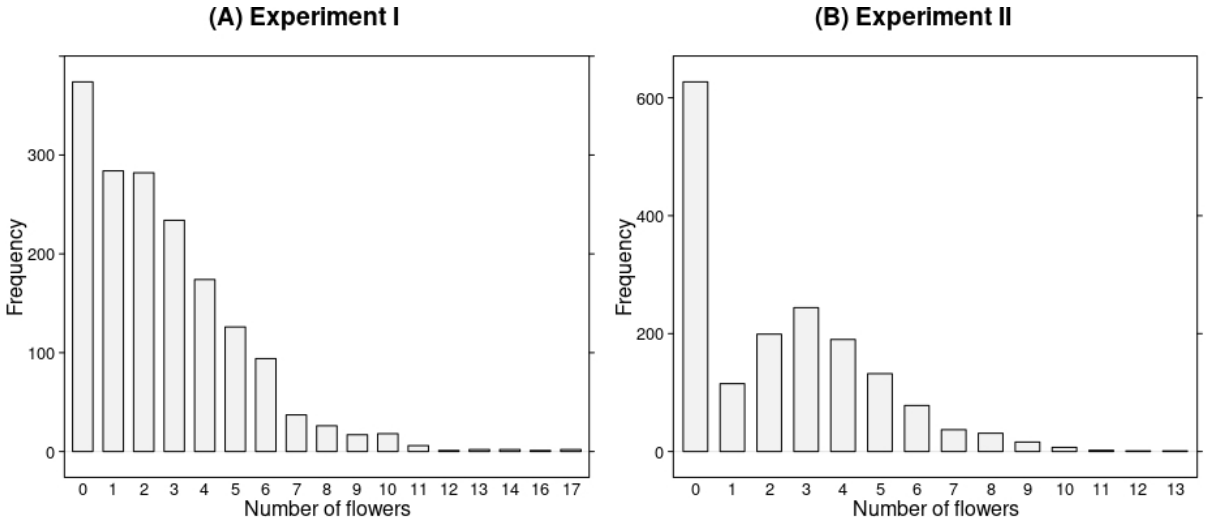

Figure 1.13. Frequency distribution for the number of flowers.

$$\eta_{ijk} = \alpha + \beta_j + \beta_{1i}\text{days}_k + \beta_{2i}\text{days}_k^2, \quad j = 1, \ldots, 10, \quad i = 1, \ldots, 28 \quad \text{and} \quad k = 1, \ldots, 6, \quad (1.31)$$

where α is the intercept, β_j is the effect of j-th block, and β_{1i} is the effect of the i-th isolate.

Looking at the analysis of deviance and goodness-of-fit given in Table 1.11, there is evidence from the residual deviance components and X^2 values that the model does not fit to the data satisfactorily, the observations are more variable than we would expect under a Poisson model. A Poisson model is clearly inadequate here with a residual deviance of 2797.6 on 1587 df indicating huge overdispersion in experiment I and a residual deviance of 2185.5 on 1587 df indicating huge overdispersion in experiment II.

This can also be seen in the half normal plot simulated envelope for the deviance residuals components shown for both the experiments in Figure 1.14 that the Poisson model does give an adequate fit to the observed values and thus it should not be used. This occurs

Table 1.11. Analysis of deviance for the number flowers data, using a Poisson log-linear model of all experiments.

Experiment I					
Sources of variation	df	Deviance	p-value	X^2	p-value
Block	9	22.72			
Isolates	27	84.92			
Days	1	821.59			
Days^2	1	23.03			
Isol:Days	27	142.73			
Isol:Days^2	27	164.25			
Residual	1587	2797.6	<0.01	2436.297	<0.01
Experiment II					
Sources of variation	df	Deviance	p-value	χ^2	p-value
Block	9	9.51			
Isolates	27	52.53			
Days	1	2138.62			
Days^2	1	257.10			
Isol:Days	27	46.58			
Isol:Days^2	27	32.60			
Residual	1587	2185.5	<0.01	2442.468	<0.01

because there is more variability than the Poisson model accommodates, it is suggested that we may try to accommodate the extra variability by estimating the dispersion parameter with a quasi-Poisson model (Demétrio et al., 2014).

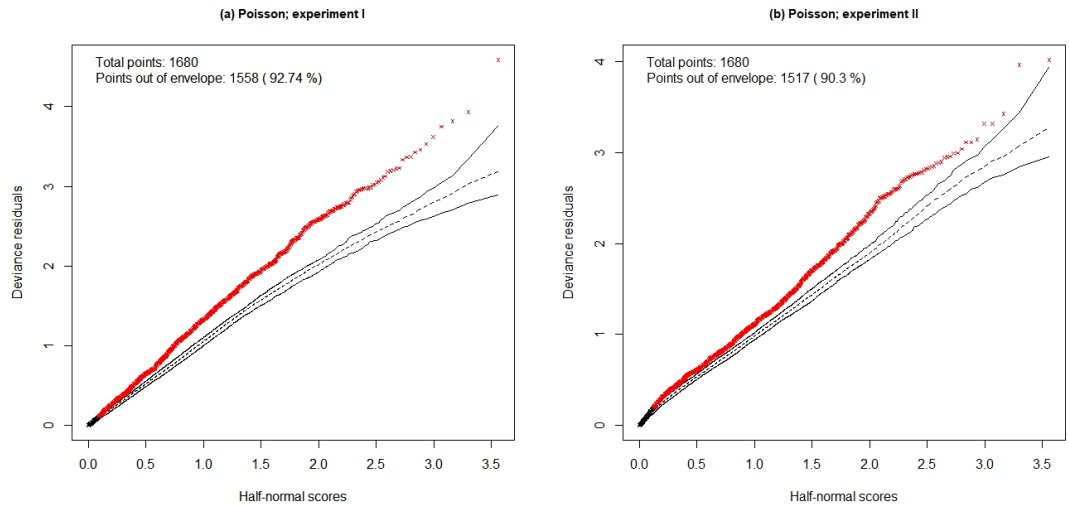

Figure 1.14. Half-normal plot with simulation envelopes of deviance residuals component using Poisson model.

1.5.3 Quasi-Poisson model

Fitting a Quasi-Poisson model with the same predictor (1.31), the estimated values of ϕ are $\tilde{\phi_1} = 1.5351$ and $\tilde{\phi_2} = 1.539$, for experiments I and II, respectively. A half-normal plot with a simulated envelope show that for experiment I, there is strong evidence of an inadequate model fit, with $23,71\%$ of the observed residuals lying outside the simulated envelope (Figure 1.15 (a)). The plot presented in (Figure 1.15 (b)) shows evidence of an adequate model, with most of the observed residuals lying inside the simulated envelope for experiment II.

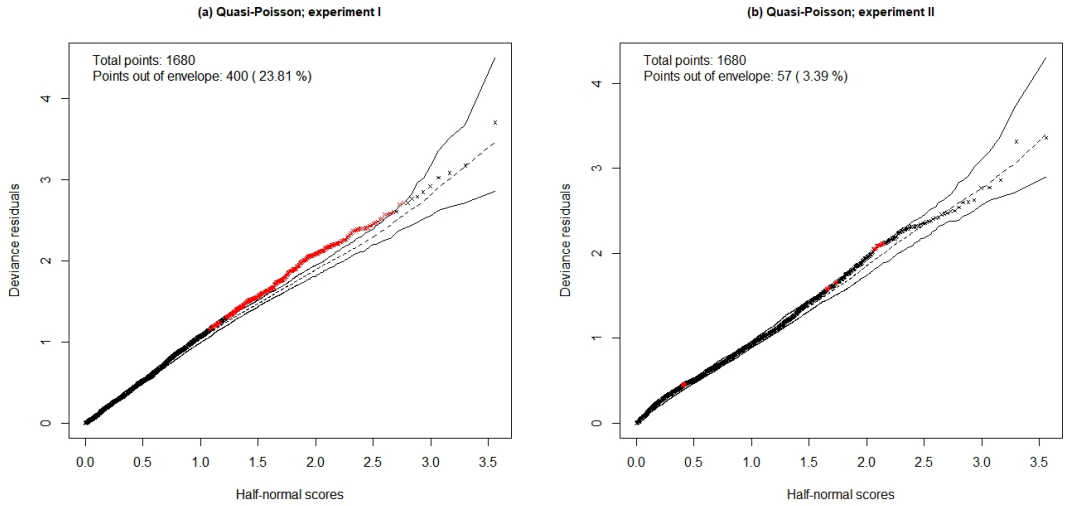

Figure 1.15. Half-normal plot with simulation envelopes of deviance residuals component using Quasi-Poisson model.

1.5.4 Negative Binomial model

The negative binomial model is an alternative approach to account for overdispersion. We can fit this model with the same linear predictor (1.31). The estimated values for θ is $\hat{\theta} = 4.86$ and $\hat{\theta} = 10.5$, for experiments I and II, respectively.

The half normal plot presented in Figure 1.16 (a) and Figure 1.16 (b) show evidence that the negative binomial model is inadequate for the data of both experiments, there is a considerable amount of points outside of the simulated envelopes.

1.5.5 Zero-inflated Poisson model

The plot of the frequency distribution for number of flowers given in Figure 1.13 shows that there are large numbers of zero observations. Alternative models to be considered are a zero inflated Poisson (ZIP) and a negative binomial models (ZINB), to incorporate excess zeros.

We, initially, fit a zero-inflated Poisson model with constant zero-inflation and with the same linear predictor (1.31) using the **R** package `pscl` Zeileis et al. (2008). A zero-inflated Poisson is clearly inadequate for experiment I when looking at the half-normal plot with most of the observed residuals lying outside the simulated envelope in Figure 1.17(a), which still suggests considerable overdispersion, while for Experiment II there is evidence of a good fit.

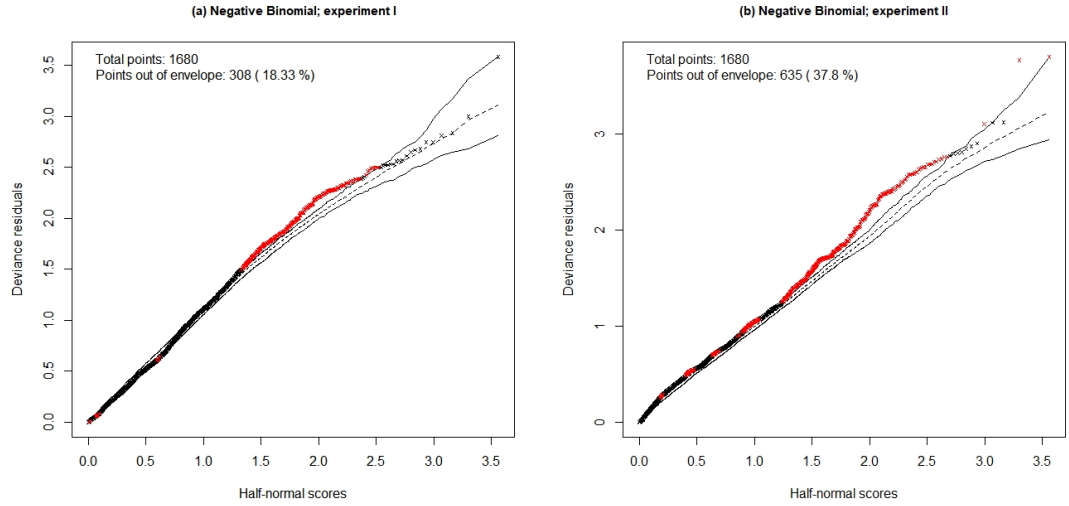

Figure 1.16. Half-normal plot with simulation envelopes of deviance residuals component using negative binomial model.

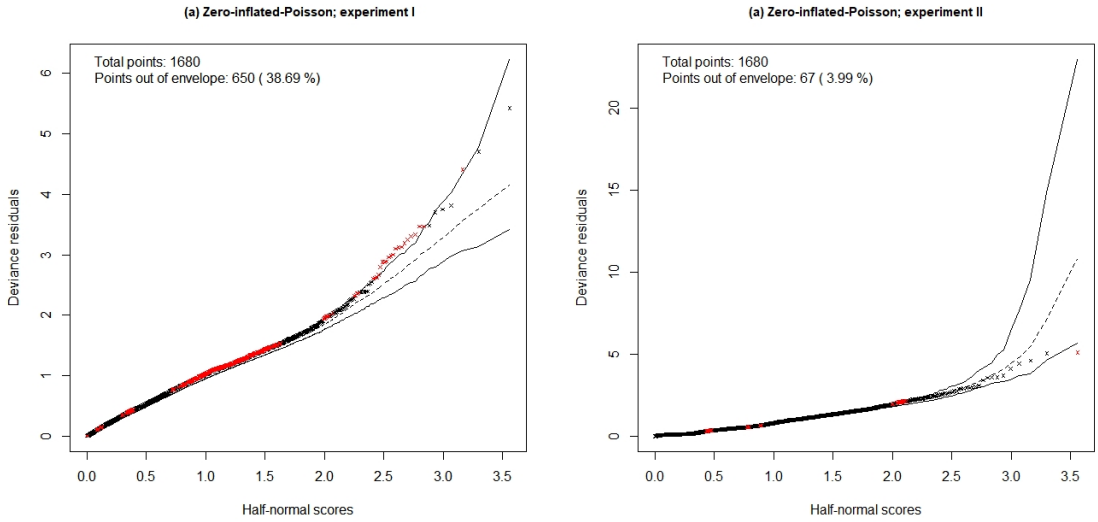

Figure 1.17. Half-normal plot with simulation envelopes of deviance residuals component using zero-inflated Poisson model.

1.5.6 Zero-inflated negative binomial model

Fitting a zero-inflated negative binomial distribution to the data, with the same linear predictor (1.31), the half normal plots presented in Figure 1.18 (a) and Figure 1.18 (b) show evidence that the zero-inflated negative binomial model is adequate for the data of both experiments, with most of the observed residuals lying inside the simulated envelope.

1.5.7 Grouping

As suggested by Fatoretto et al. (2018), the isolates can be grouped by visual inspection using a likelihood-ratio test to identify the number of groups based on the similarity observed between isolate predicted means Figure 1.19. For the analysis of number of flowers, using the

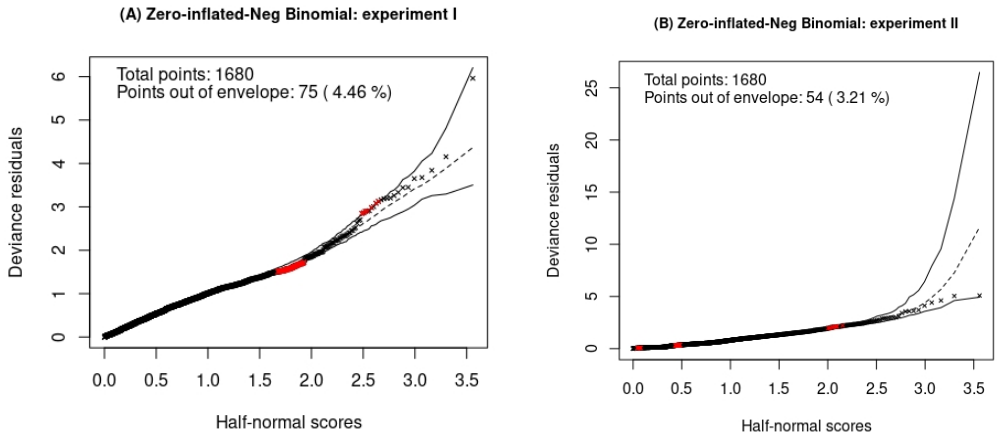

Figure 1.18. Half-normal plot with simulation envelopes of deviance residuals component using zero-inflated negative binomial model.

isolate predicted means, by visual inspection, we started creating eight groups of similar isolates for all experiment as shown in Table 1.12 and Figure 1.19. We first compared the model with 28 isolates with the model with eight groups for all experiment, resulting in a non significant likelihood ratio test. Similarities between the isolate groups were searched by merging groups, with the aim of reducing the number of groups.

In experiment I, G II was created to test similarities between the isolates 1684 and isolates of G II, but this hypothesis was not rejected. G III was created to test similarities between the isolates 1684, isolates of G II and isolates of G III, but this hypothesis was not rejected. G IV was created to test similarities between G III and G IV and this hypothesis was not rejected. G V was created to test similarities between G III, IV and V. G VI was created to test similarities between G V and G VI. G VII was created to test similarities between G V, VI and VII. And finally, G VIII was created to test similarities between G VII and VIII, but this hypothesis was not rejected. (Table 1.12).

In experiment II, G II was created to test similarities between the isolates 1629 and isolates of G II, but these hypothesis was not rejected. G III was created to test similarities between the isolates 1629, isolates of G II and isolates of G III and this hypothesis was not rejected. G IV was created to test similarities between G III and G IV, but these hypothesis was not rejected. G V was created to test similarities between G III, IV and V, and this hypothesis was not rejected. G VI was created to test similarities between G V and G VI. G VII was created to test similarities between G V, VI and VII. And finally G VIII was created to test similarities between G VII and VIII, but this hypothesis was rejected (Table 1.12).

According to the tests, we can group the isolates in four groups in experiments I and experiment II (Table 1.13) in which the isolates belonging to distinct groups are significantly different at a significance level of 5%. The four groups are isolates that give variable number of flowers

- group I: isolates with almost no effect (similar to the worst control) with smaller number of flowers;

Table 1.12. First grouping according to the empirical grouping of the isolate predicted means.

Groups	Experiment I				
G I	1684				
G II	1451	1638			
G III	3692	1636	Quar		
G IV	1637	PL63	1634	3693	1669
	1587	1618	1604		
G V	1641	1610	1608	3323	Cont
	1296	3703	1622		
G VI	1709				
G VII	1629	3375			
G VIII	43	1306	1635		
Groups	Experiment II				
G I	1629				
G II	1610	1669	1306	3692	
G III	3703	1618	3693	Quar	3323
G IV	Cont	1637	1634	1636	43
G V	1622	1587	1635	1684	1296
	PL63	1604	1638		
G VI	3375	1709			
G VII	1608	1641			
G VIII	1451				

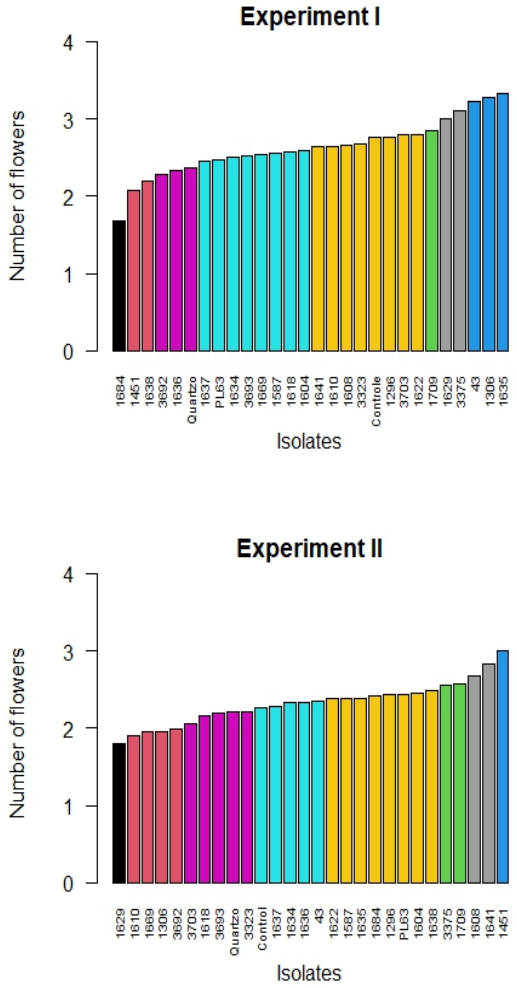

Figure 1.19. Plot with the predicted average values

- group II: moderately promising isolates;
- group IV: less highly promising isolates;
- group IV: highly promissing isolates (larger number of flowers).

1.5.8 Discussion

In this Section, we proposed different models that take into account overdispersion and zero inflation, to analyse the number of flowers after root inoculation of strawberry plants inoculated with different promising isolates of the entomopathogenic fungi of *Metarhizium spp., B. bassiana, I. fumosorosea*. We compared the results and also discussed model selection and diagnostics. For grouping the isolates we proposed one empirical method. The methods were implemented in the software R (R Core Team, 2020) and the scripts developed are presented in the Appendix II.

Table 1.13. Final grouping according to the empirical grouping of the isolate predicted means.

Groups	Experiment I				
G I	1684				
G II	1451	1638	3692	1636	Quar
	1637	PL63	1634	3693	1669
	1587	1618	1604	1641	1610
	1608	3323	Cont	1296	3703
	1622	1709			
G III	1629	3375			
G IV	43	1306	1635		
Groups	Experiment II				
G I	1629	1610	1669	1306	3692
	3703				
G II	1618	3693	Quar	3323	Cont
	1637	1634	1636	43	
G III	1622	1587	1635	1684	1296
	PL63	1604	1638	3375	1709
G IV	1608	1641	1451		

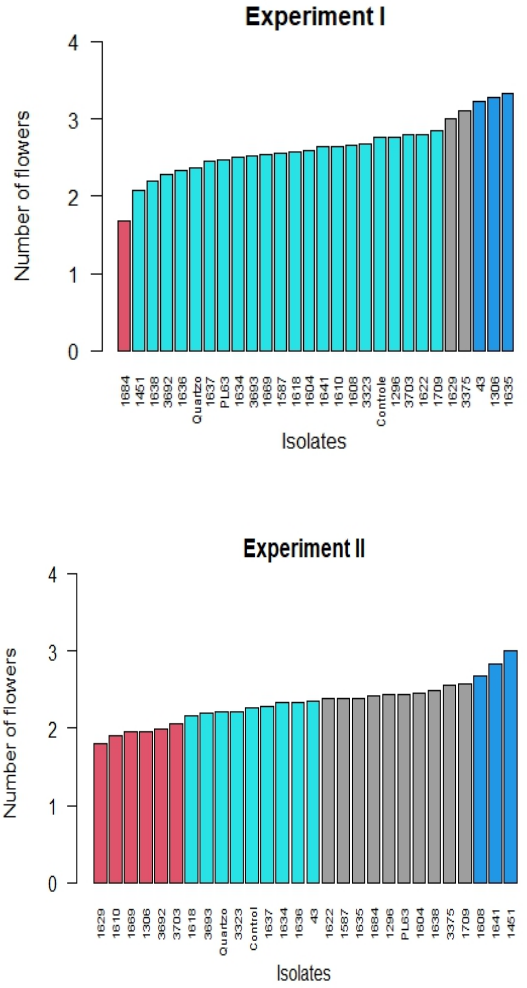

Figure 1.20. Plot with the predicted average values

1.6 Analysis the case-study - number of leaves

The dispersion plots of the number of number of leaves for each isolates, over time for experiments I and II Figures 1.21 (a) show an increasing trend over time and that there are clear differences in the influence of the isolates and evidence of differing degrees of variability between replicates.

The dispersion plots of the sample variance versus sample means Figures 1.21 (b) show that there are points below and above the identity line, suggesting evidence of overdispersion and underdispersion.

1.6.1 Poisson model

We begin by fitting a Poisson log-linear model with the factors block, isolate, and day as fixed effects, using the maximal linear predictor given by the equation (1.32).

$$\eta_{ijk} = \alpha + \beta_j + \beta_{1i}\text{days}_k + \beta_{2i}\text{days}_k^2, \quad j = 1, \ldots, 10, \quad i = 1, \ldots, 28 \quad \text{and} \quad k = 1, \ldots, 7, \quad (1.32)$$

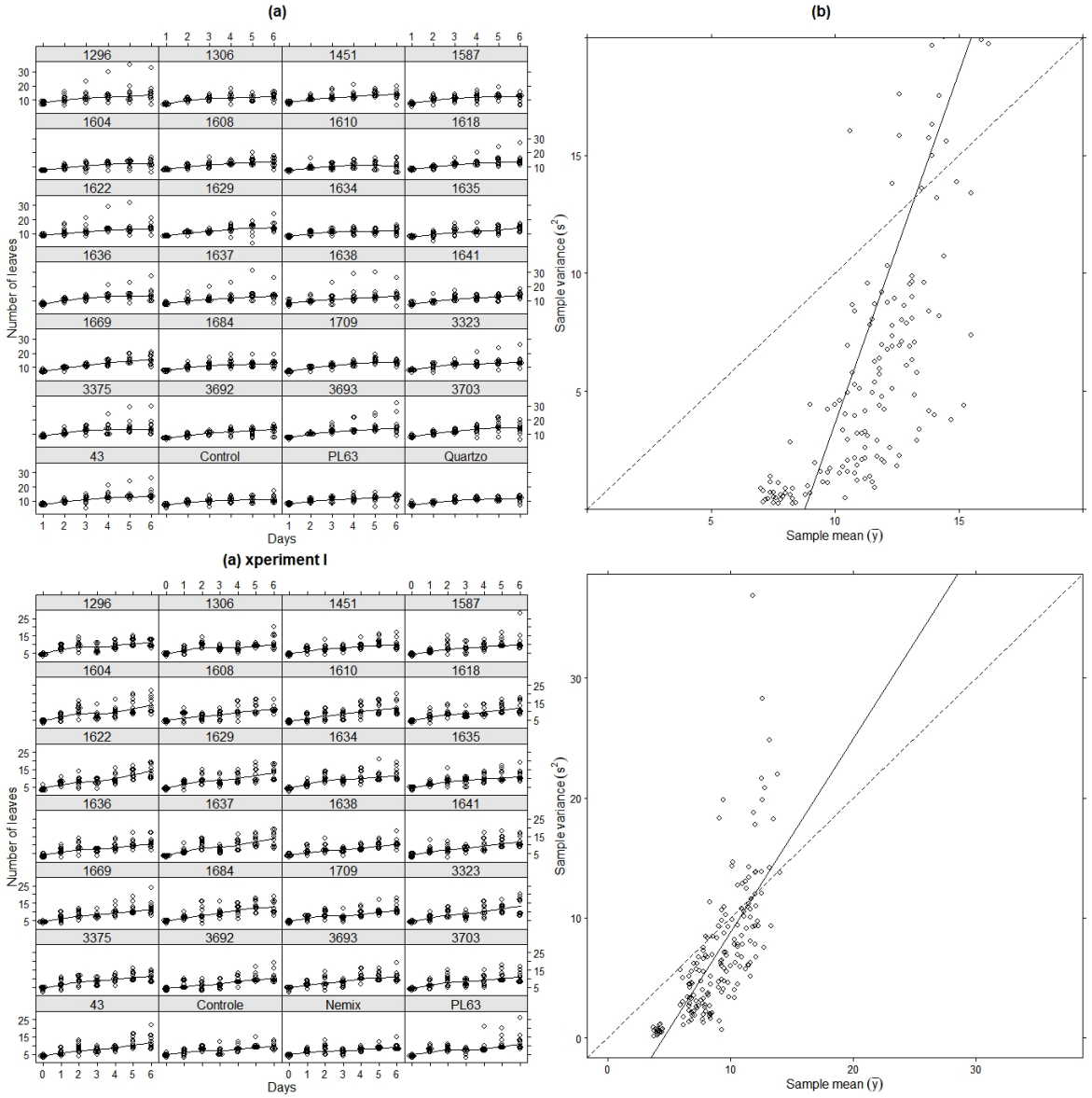

Figure 1.21. Dispersion plots of (a) number of flowers per day versus days and (b) sample mean versus the sample variance of number of leaves per day (dotted line is the identity line and solid line least squares line.

where α is the intercept, β_j is the effect of j-th block, and β_{1i} is the effect of the i-th isolate.

Looking at the analysis of deviance and goodness-of-fit given in Table 1.14, there is evidence from the residual deviance components and X^2 values that the model does not fit the data satisfactorily, the observation are more variable than we would expect under a Poisson model.

This can also be seen in the half normal plot simulated envelope for the deviance residuals components shown for both the experiments in Figure 1.22 (A) that the Poisson model does not give an adequate fit to the observed values and thus it should not be used. This occurs because there is more variability than the Poisson model accommodates, it is suggested that we may try to accommodate the extra variability by estimating the dispersion parameter with a quasi-Poisson model (Demétrio et al., 2014).

Table 1.14. Analysis of deviance for the number of leaves data, using a Poisson log-linear model.

Experiment I					
Sources of variation	df	Deviance	p-value	X^2	p-value
Block	9	11.86			
Isolates	27	63.43			
Days	1	1270.96			
Days^2	1	68.45			
Residual	1921	1422.30	<0.01	1483.84	<0.01
Experiment II					
Sources of variation	df	Deviance	p-value	X^2	p-value
Block	9	30.36			
Isolates	27	98.76			
Days	1	543.24			
Days^2	1	92.67			
Residual	1641	918.09	<0.01	968.51	<0.01

1.6.2 Quasi-Poisson model

Fitting a Quasi-Poisson model with the same predictor (1.32) the estimated values of ϕ are $\tilde{\phi_1} = 0.77$ and $\tilde{\phi_2} = 0.59$, for experiments I and II, respectively.

A half-normal plot with a simulated envelope Figure 1.22 (B) show that for experiments I and II, there is strong evidence of an inadequate model fit, with most of the observed residuals lying outside the simulated envelope which shows evidence of an inadequate model (Demétrio et al., 2014).

1.6.3 Negative Binomial model

Fitting the negative binomial model with the same liner predict (1.32). The estimated values for θ is $\hat{\theta_1} = 154242.4$ and $\hat{\theta_2} = 351007$ for experiments I and II, respectively.

The half normal plots presented in Figure 1.22 (C) show evidence that the negative binomial model is inadequate for analysing this set of data.

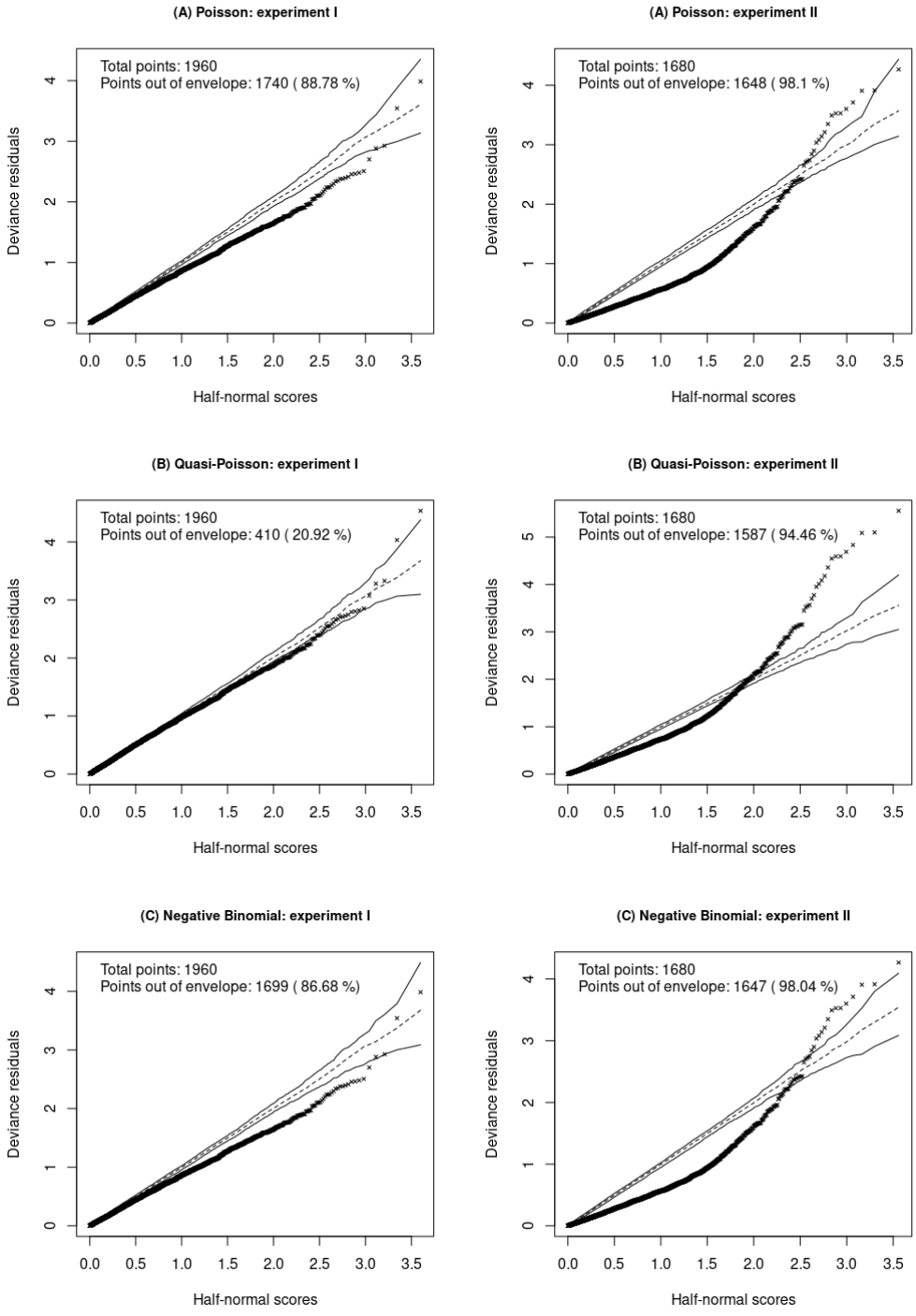

Figure 1.22. Half-normal plot with simulation envelopes for deviance residuals for (A) Poisson, (B) quasi-Poisson and (C) Negative binomial, for the number of leaves.

1.6.4 COM-Poisson model

Alternatives analysis of data with underdispersion and overdispersion have been proposed by Huang (2017) and Ribeiro Jr et al. (2020), by using different mean parametrizations of the COM-Poisson model.

We fitted a COM-Poisson model (in the two forms (original and new parametrization)

with the same predictor (1.32) for the mean ($\eta_{ijk} = \log(\mu_{ijk})$).

The parameter estimated values of ϕ and goodness-of-fit measures (log-likelihood, AIC and BIC) for the Poisson, COM-Poisson, COM-Poisson$_\mu$ and Quasi-Poisson are given in Table 1.15.

The results presented in Table 1.15 show that the goodness-of-fit measures are quite similar for the COM-Poisson and COM-Poisson$_\mu$ models. In line with Ribeiro Jr et al. (2020) the reparametrization does not change the model fit.

The Poisson model is unsuitable, being conservative, due to the overestimated standard errors. The -2 x difference between the log-likelihood of the Poisson and COM-Poisson$_\mu$ model was 190.522 for one additional parameter, which confirms the significantly fit of the COM-Poisson model. The estimated value of the dispersion parameter $\hat{\phi} = 0.530$ indicates underdispersion. Another possible model under study is the COM-Poisson with varying dispersion.

1.6.5 Discussion

In this Section, we proposed different models that take into account overdispersion (underdispersion) to analyse the number of leaves after root inoculation of strawberry plants inoculated with different promising isolates of the entomopathogenic fungi of Metarhizium spp., B. bassiana, I. fumosorosea. We compared the results and also discussed model selection and diagnostics, but the difficulty in programming. The database is difficult to analyzed. We are having difficulties in developing adapted methodologies due to the nature of the data. All the methods were implemented in the software R (R Core Team, 2020) and the scripts developed are presented in the Appendix. For the number of leaves, many other modelswere fitted with no success. Additional models need to be developed. Also, it is under development the half-normal plot for the COM-Poisson model.

1.7 Final remarks

Outcomes of interest for entomological data are often in the form of counts and as a first step, a standard model to analyse this type of data is the Poisson model, an example of generalized linear models. The basic model assumptions are independence of observations and constant rate of event occurrence. If one or both of these assumptions failure the variance of the data will be greater (smaller) than the variance expected using the Poisson model resulting in what is called overdispersion (undersispersion). Many different models for overdispersion (underdispersion) can arise from alternative possible mechanisms for the underlying process. Another reason for extending the Poisson model is because of the occurrence of a hierarchical structure in the data caused by a clustering resulted from repeatedly measuring the outcome on the same experimental unit. In entomological applications involving count data there is often an excess of zero observations. In this work we present a review of models that can be used to take into account the different aspects of the failure of the Poisson model assumptions. The proposed methodology is illustrated using data of an experiment to evaluate 25 isolates of entomopathogenic fungi (*Metarhizium spp., B. bassiana and I. fumosorosea*) and compare with

Table 1.15. Parameter estimates (Est) and standard error (SE) for the five model for the analysis of the experiment .

Experiment				
Parameter	Poisson Est(SE)	COM-Poisson Est(SE)	CMP$_\mu$ Est(SE)	Quasi-Poisson Est(SE)
ϕ,σ		0.530 (15.086)	0.530 (15.081)	0.648
Intercept	2.220 (49.339)	3.810 (25.796)	2.220 (63.796)	2.220 (61.306)
trt1306	-0.149 (-2.775)	-0.248 (-3.558)	-0.149 (-3.587)	-0.149 (-3.448)
trt1451	-0.027 (-0.519)	-0.045 (-0.670)	-0.027 (-0.673)	-0.027 (-0.645)
trt1587	-0.139 (-2.610)	-0.233 (-3.348)	-0.140 (-3.374)	-0.139 (-3.243)
trt1604	-0.163 (-3.024)	-0.271 (-3.872)	-0.163 (-3.909)	-0.163 (-3.758)
trt1608	-0.070 (-1.337)	-0.117 (-1.723)	-0.070 (-1.730)	-0.070 (-1.661)
trt1610	-0.201 (-3.696)	-0.335 (-4.713)	-0.201 (-4.779)	-0.201 (-4.593)
trt1618	-0.037 (-0.702)	-0.061 (-0.906)	-0.036 (-0.907)	-0.037 (-0.872)
trt1622	0.013 (0.257)	0.022 (0.331)	0.013 (0.331)	0.013 (0.319)
trt1629	-0.048 (-0.912)	-0.079 (-1.177)	-0.048 (-1.180)	-0.048 (-1.134)
trt1634	-0.142 (-2.665)	-0.238 (-3.418)	-0.143 (-3.446)	-0.142 (-3.311)
trt1635	-0.108 (-2.036)	-0.180 (-2.621)	-0.108 (-2.632)	-0.108 (-2.530)
trt1636	-0.050 (-0.965)	-0.084 (-1.246)	-0.050 (-1.247)	-0.050 (-1.199)
trt1637	-0.052 (-0.992)	-0.086 (-1.279)	-0.052 (-1.281)	-0.052 (-1.232)
trt1638	-0.048 (-0.912)	-0.079 (-1.177)	-0.048 (-1.178)	-0.048 (-1.134)
trt1641	-0.108 (-2.036)	-0.180 (-2.620)	-0.108 (-2.632)	-0.108 (-2.530)
trt1669	-0.034 (-0.650)	-0.056 (-0.838)	-0.034 (-0.836)	-0.034 (-0.807)
trt1684	-0.092 (-1.739)	-0.153 (-2.239)	-0.092 (-2.248)	-0.092 (-2.160)
trt1709	-0.086 (-1.631)	-0.143 (-2.100)	-0.086 (-2.106)	-0.086 (-2.027)
trt3323	-0.037 (-0.702)	-0.061 (-0.906)	-0.037 (-0.908)	-0.037 (-0.872)
trt3375	0.035 (0.690)	0.059 (0.892)	0.035 (0.892)	0.035 (0.857)
trt3692	-0.175 (-3.247)	-0.292 (-4.152)	-0.175 (-4.194)	-0.175 (-4.035)
trt3693	0.031 (0.614)	0.053 (0.794)	0.032 (0.797)	0.031 (0.762)
trt3703	-0.013 (-0.259)	-0.022 (-0.333)	-0.013 (-0.334)	-0.013 (-0.321)
trt43	-0.064 (-1.230)	-0.108 (-1.587)	-0.064 (-1.586)	-0.064 (-1.529)
trtControl	-0.230 (-4.208)	-0.384 (-5.342)	-0.231 (-5.435)	-0.230 (-5.228)
trtPL63	-0.146 (-2.720)	-0.243 (-3.488)	-0.146 (-3.516)	-0.146 (-3.379)
trtQuartzo	-0.204 (-3.753)	-0.340 (-4.781)	-0.204 (-4.849)	-0.204 (-4.663)
days	0.098 (23.204)	0.164 (20.919)	0.098 (29.944)	0.098 (28.832)
LogLik	-4084.046	-3988.555	-3988.785	-
AIC	8244.092	8055.109	8055.569	-
BIC	8450.301	8266.745	8267.205	-

the three reference treatments on the control of *T. urticae*. We compared the results and also discussed model selection and diagnostics. For grouping the isolates we proposed two different methods. All the methods were implemented in the software `R`.

ACKNOWLEDGMENT

The first author acknowledges the scholarship financial support from Commissão de Aperfeicoamento de Pessoal de Níivel Superior (CAPES), Brazil. This research was partially supported by FAPESP, Brazil CNPq for CGBD.

Referências

Akaike, H. (1973). Information theory and an extention of the maximum likelihood principle. In *2nd International Symposium on Information Theory, 1973*, pages 267–281. Akademiai Kiado.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 33(1):38–44.

Canassa, F., D'Alessandro, C. P., Sousa, S. B., Demétrio, C. G., Meyling, N. V., Klingen, I., e Delalibera Jr, I. (2020). Fungal isolate and crop cultivar influence the beneficial effects of root inoculation with entomopathogenic fungi in strawberry. *Pest management science*, 76(4):1472–1482.

Castro, T. R. d. (2011). Estudos para o desenvolvimento de metodologia da produção in vivo do fungo neozygites floridana weiser e muma para controle do ácaro tetrancychus urticae koch. Master's thesis, Universidade de São Paulo.

Ceuppens, S., Johannessen, G., Allende, A., Tondo, E., El-Tahan, F., Sampers, I., Jacxsens, L., e Uyttendaele, M. (2015). Risk factors for salmonella, shiga toxin-producing escherichia coli and campylobacter occurrence in primary production of leafy greens and strawberries. *International journal of environmental research and public health*, 12(8):9809–9831.

Conway, R. W. e Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136.

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

Fahim, A., Salem, A., Torkey, F. A., e Ramadan, M. (2006). An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University-Science A*, 7(10):1626–1633.

Fatoretto, M. B., Moral, R. d. A., Demétrio, C. G. B., de Pádua, C. S., Menarin, V., Rojas, V. M. A., D'Alessandro, C. P., e Delalibera Jr, I. (2018). Overdispersed fungus germination data: statistical analysis using r. *Biocontrol Science and Technology*, 28(11):1034–1053.

Gbur, E. E., Stroup, W. W., McCarter, K. S., Durham, S., Young, L. J., Christman, M., West, M., e Kramer, M. (2012). *Generalized linear mixed models*. American Society of Agronomy, Madison, WI.

Hartigan, J. A. e Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

Hinde, J. (1982). Compound poisson regression models. In *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, pages 109–121. Springer.

Hinde, J. e Demétrio, C. G. B. (1998). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.

Huang, A. (2017). Mean-parametrized conway–maxwell–poisson regression models for dispersed counts. *Statistical Modelling*, 17(6):359–380.

Kovaleski, A. K., Bortolozzo, A. R. B., Hoffmann, A. H., Calegario, F. F. C., MELO, G. W. B. D. M., Bernardi, J. B., Vargas, L. V., Botton, M. B., Ferla, N. J. F., e Pinent, S. M. J. P. (2006). *Produção de morangos no sistema semi-hidropônico*. Embrapa Uva e Vinho.

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

Lindsey, J. (1995). Fitting parametric counting processes by using log-linear models. *J R Stat Soc Ser C Appl Stat*, 44:201–212.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

McCullagh, P. e Nelder, J. A. (1989). Generalized linear models 2nd edition chapman and hall. *London, UK*.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2017). Hierarchical models with normal and conjugate random effects: a review. *SORT*, 41:191:254.

Molenberghs, G., Verbeke, G., Demétrio, C. G., Vieira, A. M., et al. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, 25(3):325–347.

Moral, R. A., Hinde, J., e Demétrio, C. G. (2017). Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 81:1–23.

Nazeer, K. A. e Sebastian, M. (2009). Improving the accuracy and efficiency of the k-means clustering algorithm. In *Proceedings of the world congress on engineering*, volume 1, pages 1–3. Citeseer.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr, E. E., Demétrio, C. G. B., e Hinde, J. (11/07/2019). Double com-poisson models with varying dispersion. In *International Workshop on Statistical Modelling, 34. Proceedings of the* 34^{th} *International Workshop on Statistical Modelling*, pages 1:101–106, Guimarães, Portugal. Printed by Instituto Nacional de Estatística, ISBN 978-989-20-9528-8.

Ribeiro Jr, E. E., Zeviani, W. M., Bonat, W. H., Demetrio, C. G., e Hinde, J. (2020). Reparametrization of com-poisson regression models with applications in the analysis of experimental data. *Statistical Modelling*, 20:443–466.

Ridout, M., Demétrio, C. G., e Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192. International Biometric Society Invited Papers Cape Town, South Africa.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

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

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., e Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Sjulin, T. M. (2003). The north american small fruit industry 1903-2003. ii. contributions of public and private research in the past 25 years and a view to the future. *HortScience*, 38(5):960–967.

Verbeeke, G. e Molenberghs, G. (2000). Linear mixed models for longitudinal data.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447.

Zeileis, A., Kleiber, C., e Jackman, S. (2008). Regression models for count data in r. *Journal of statistical software*, 27(8):1–25.

APÊNDICES

Apêndice I: computational routines

We carried out all programming in R For Poisson, Quasi-Poisson, negative binomial, Poisson-normal, Negative-binimial-normal and COM-Poisson models the following codes were used for number of eggs for experiment I:

```
source("helper01_general-functions.R")
source("helper02_lattice-panels.R")

# Predictor
f1 <- neggs ~ block + isol

# Poisson fit complet
model1 <- glm(f1, family=poisson, data = dados)
anova(model1, test="Chisq")
sum(resid(model1, ty="pearson")^2)
summary(model1)
logLik(model1)
hnp(model1, print.on = T, pch=4, main="(A) Poisson: experiment I",
    cex=0.5, cex.main=0.9, pty='s', xlab="Half-normal scores",
    lab="Deviance residuals")

# Quasi-Poisson fit
model2 <- glm(f1, family=quasipoisson, data = dados)
summary(model2)$dispersion #phi
summary(model2)
logLik(model2)
anova(model2, test="F")
hnp(model2, print.on = T, pch=4, main="(B) Quasi-Poisson:
experiment I", cex=0.5, cex.main=0.9, pty='s',
xlab="Half-normal scores", ylab="Deviance residuals")

# Negative binomial fit
model3 <- glm.nb(f1, data=dados)
thetahat <- summary(model3)$theta #theta estimat
anova(model3, test = "F")
hnp(model3, print.on = T, pch=4, main="(C) Negative binomial:
```

```
experiment␣I", cex=0.5, cex.main=0.9, pty='s',
xlab="Half-normal␣scores", ylab="Deviance␣residuals")

# Poisson-normal model
z <- factor(c(rep(1:28, each=5)))
id <- factor(1:nrow(numbereggs))
modelPN <- glmer(NO ~ block + trt  + (1 | id),
                 family = poisson, data = numbereggs)
summary(modelPN)
anova(modelPN, test = "F")
logLik(modelPN)
getME(modelPN, "theta")^2 # Normal variance parameter
hnp(modelPN, paint.out = T, print.on = T)
#-------------------------------------------------------
modelPN1 <- glmer(NO ~ block + trt + (1 | z),
                 family = poisson, data = numbereggs)
logLik(modelPN1)
hnp(modelPN1, paint.out = T, print.on = T)
#-------------------------------------------------------
modelPN2 <- glmer(NO ~ block + (1 | id) + (1|z),
                  family = poisson, data = numbereggs)
summary(modelPN2)
anova(modelPN2, test = "F")
logLik(modelPN2)
getME(modelPN2, "theta")^2 # Normal variance parameter
hnp(modelPN2, paint.out = T, print.on = T)

#-------------------------------------------------------------
# Negative-binomial-normal model (combined approach)
modelCB <- glmer.nb(NO ~ block + trt + (1 | id), data = numbereggs,
                    control=glmerControl(optimizer = 'bobyqa',
                     optCtrl=list(maxfun=600000)))
summary(modelCB)
logLik(modelCB)

getME(modelCB, "glmer.nb.theta") # Negative-binomial parameter
getME(modelCB, "theta")^2        # Normal variance parameter

# Implementando hnp para o modelo binomial Negativo.
```

```
# Variável resposta:neggs
resp<-numbereggs$NO

dfun <- function(obj) resid(obj,type="deviance")

sfun <- function(n, obj) simulate(obj)[[1]]

ffun <- function(resp)glmer.nb(resp~ block + trt+ (1|id),
data=numbereggs,control=glmerControl(optimizer= 'bobyqa'
,optCtrl=list(maxfun=600000)))


#hnp
set.seed(1618)
hnp(modelCB , conf = 0.95,newclass = TRUE, verb.sim = T,
    diagfun = dfun, simfun = sfun,
    fitfun = ffun,print = TRUE, print.on = T,
    pch=4, main="(E) Negative-binomial-normal: experiment I",
    cex=.5, cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")


#-----------------------------------------------------------------
modelCB1 <- glmer.nb(NO ~ block + trt + (1 | z), data = numbereggs,
                     control=glmerControl(optimizer = 'bobyqa',
                   optCtrl=list(maxfun=600000)))
summary(modelCB1)
logLik(modelCB1)

getME(modelCB1, "glmer.nb.theta") # Negative-binomial parameter
getME(modelCB1, "theta")^2        # Normal variance parameter

# Implementando hnp para o modelo binomial Negativo.
# Variável resposta:neggs
resp<-numbereggs$NO

dfun <- function(obj) resid(obj,type="deviance")

sfun <- function(n, obj) simulate(obj)[[1]]
```

```
ffun <- function(resp)glmer.nb(resp~ block + trt+ (1|z),
data=numbereggs, control=glmerControl(optimizer= 'bobyqa'
,optCtrl=list(maxfun=600000)))


#hnp
set.seed(1618)
hnp(modelCB1 , conf = 0.95,newclass = TRUE, verb.sim = T,
    diagfun = dfun, simfun = sfun, fitfun = ffun,print = TRUE)


#-----------------------------------------------------------------
modelCB2 <- glmer.nb(NO ~ block + trt + (1 | id) + (1 | z),
data = numbereggs)
summary(modelCB2)
logLik(modelCB2)

getME(modelCB2, "glmer.nb.theta") # Negative-binomial parameter
getME(modelCB2, "theta")^2        # Normal variance parameter

resp<-numbereggs$NO

dfun <- function(obj) resid(obj,type="deviance")

sfun <- function(n, obj) simulate(obj)[[1]]

ffun <- function(resp)glmer.nb(resp~ block + trt + (1|z) + (1 | id),
data=numbereggs, control=glmerControl(optimizer= 'bobyqa'
,optCtrl=list(maxfun=600000)))

#hnp
set.seed(1618)
hnp(modelCB2 , conf = 0.95,newclass = TRUE, verb.sim = T,
    diagfun = dfun, simfun = sfun, fitfun = ffun,print = TRUE,
    print.on = T, pch=4, main="(E) Negative-binomial-normal:
    experiment I", cex=.5, cex.main=0.9, pty='s',
    lab="Half-normal scores",      ylab="Deviance residuals")
```

```
#-------------------------------------------------------------------------
#Empirical grouping - Visual inspectio
predito <- predict(model2, type = "response")
media <- tapply(predito, dados$isol, mean)
med.ord <- sort(media)
par(mar = c(6.5, 6.5, 2, 1) + 0.1)
bpp <- barplot(med.ord, beside=TRUE, border = "black",
                ylim = c(0,60),
                col = "lightgray",
                xlab = "isolates",
                ylab = "Predicted values of number of eggs  ",
                main = "Experiment I", las=2)


#-------------------------------------------------------------------------
#Creating the factor grouping 1
dados$grouping1 <- dados$isol
levels(dados$grouping1)
levels(dados$grouping1) <- c(8,9,6,6,4,7,4,6,1,8,6,6,4,7,5,2,
6,5,6,6,6,4,6,7,3,10,8,9)

#model fitted with factor grouping 1 instead of fung
model5 <- glm(neggs ~ block + grouping1 , family=quasipoisson,
data = dados)

#-------------------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model5, model2, test = "F")

# The model5 is selected, because p>0.05 , the isolates within
#the groups do not differ statistically.
#-------------------------------------------------------------------------
#Creating the factor grouping 2
dados$grouping2 <- dados$isol
levels(dados$grouping2)
levels(dados$grouping2) <- c(8,9,6,6,4,7,4,6,1,8,6,6,4,7,
5,1,6,5,6,6,6,4,6,7,3,10,8,9)
```

```
#model fitted with factor grouping 2 instead of fung
model6 <- glm(neggs ~ block + grouping2 , family=quasipoisson,
data = dados)

#-----------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model6, model2, test = "F")
anova(model6, model5, test = "F")

# The model6 is selected, because p>0.05 ,
the isolates contrel does not differ #from isolates in group2
#-----------------------------------------------------------------
#Creating the factor grouping 3
dados$grouping3 <- dados$isol
levels(dados$grouping3)
levels(dados$grouping3) <- c(8,9,6,6,4,7,4,6,1,8,6,6,4,7,5,1
,6,5,6,6,6,4,6,7,1,10,8,9)

#model fitted with factor grouping 3 instead of fung
model7 <- glm(neggs ~ block + grouping3, family=quasipoisson,
data = dados)

#-----------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model7, model2, test = "F")
anova(model7, model5, test = "F")
anova(model7, model6, test = "F")

# The model6 is selected, because p<0.05 , the isolates
control, Nemix and 1306 #differs from isoletes in group3.
#-----------------------------------------------------------------
#Creating the factor grouping 4
dados$grouping4 <- dados$isol
levels(dados$grouping4)
levels(dados$grouping4) <- c(8,9,6,6,4,7,4,6,1,8,6,6,4,7,4,
```

```
1,6,4,6,6,6,4,6,7,1,10,8,9)


#model fitted with factor grouping 4 instead of fung
model8 <- glm(neggs ~ block + grouping4, family=quasipoisson,
data = dados)

#-----------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT

anova(model8, model2, test = "F")
anova(model8, model5, test = "F")
anova(model8, model6, test = "F")
anova(model8, model7, test = "F")

# The model8 is selected, because p>0.05 , the isolates 1296,
PL63 and 1629 #does not differs from isoletes in group2
#-----------------------------------------------------------------
#Creating the factor grouping 5
dados$grouping5 <- dados$isol
levels(dados$grouping5)
levels(dados$grouping5) <- c(7,9,6,6,4,7,4,6,1,7,6,6,4,6,4,1,
4,4,6,6,6,4,6,7,1,10,7,9)


#model fitted with factor grouping 5 instead of fung
model9 <- glm(neggs ~ block + grouping5, family=quasipoisson,
data = dados)

#-----------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model9, model2, test = "F")
anova(model9, model5, test = "F")
anova(model9, model6, test = "F")
anova(model9, model7, test = "F")
anova(model9, model7, test = "F")
anova(model9, model8, test = "F")
```

```
# The model8 is selected, because p>0.05 , the isolates
1in group5 # does not differs from isoletes in group4
#-------------------------------------------------------------------------
#Creating the factor grouping 6
dados$grouping6 <- dados$isol
levels(dados$grouping6)
levels(dados$grouping6) <- c(7,9,6,6,4,7,4,6,1,7,6,6,4,6,4,1,
4,4,6,6,6,4,6,7,1,9,7,9)

#model fitted with factor grouping 6 instead of fung
model10 <- glm(neggs ~ block + grouping6, family=quasipoisson,
data = dados)

#-------------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model10, model2, test = "F")
anova(model10, model5, test = "F")
anova(model10, model6, test = "F")
anova(model10, model7, test = "F")
anova(model10, model8, test = "F")
anova(model10, model9, test = "F")

# The model10 is selected, because p>0.05 , the isolates in group4
#  differs from isoletes in group6
#-------------------------------------------------------------------------
#Creating the factor grouping 7
dados$grouping7 <- dados$isol
levels(dados$grouping7)
levels(dados$grouping7)<- c(7,9,6,6,4,7,4,6,4,7,6,6,4,6,4,4,
4,4,6,6,6,4,6,7,4,9,7,9)

#model fitted with factor grouping 7 instead of fung
model11 <- glm(neggs ~ block + grouping7, family=quasipoisson,
data = dados)

#-------------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
```

```
# LRT
anova(model11, model2, test = "F")
anova(model11, model5, test = "F")
anova(model11, model6, test = "F")
anova(model11, model7, test = "F")
anova(model11, model8, test = "F")
anova(model11, model9, test = "F")
anova(model11, model10, test = "F")

# The model10 is selected, because p<0.05 , the isolates in group6
#  differs from isoletes in group7
#-------------------------------------------------------------------
#-------------------------------------------------------------------
#Creating the factor grouping 8
dados$grouping8 <- dados$isol
levels(dados$grouping8)
levels(dados$grouping8) <- c(7,9,6,6,6,7,6,6,6,7,6,6,6,6,6,
6,6,6,6,6,6,6,7,6,9,7,9)

#model fitted with factor grouping 6 instead of fung
model12 <- glm(neggs ~ block + grouping8, family=quasipoisson,
data = dados)

#-------------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model12, model2, test = "F")
anova(model12, model5, test = "F")
anova(model12, model6, test = "F")
anova(model12, model7, test = "F")
anova(model12, model8, test = "F")
anova(model12, model9, test = "F")
anova(model12, model10, test = "F")
anova(model12, model11, test = "F")
# The model10 is selected, because p>0.05 , the isolates in group4
# does not differs from isoletes in group6
#-------------------------------------------------------------------
hnp(model11, print.on = T, pch=4, main="(A) Quasi-Poisson: experiment I
    cex=.5, cex.main=0.9, pty='s', xlab="Half-normal scores", ylab="Dev
```

```
library(emmeans)
medias <- emmeans::emmeans(model2, ~ isol,
                           type="response")
medias <- data.frame(medias)
colnames(medias)[2] <-"media"
medias[1:3]

#Grafico simples dos efeitos
library(effects)
effects::effect("isol", model2)
#plot(effects::effect("isol", model2))
#####
library(ggplot2)
#10grup
ggplot(medias,aes(x=reorder(isol,media), y=media))+
  geom_col(fill= c(11,9,3,3,5,8,5,3,1,11,
                   3,3,5,8,7,2,3,7,3,3,
                   3,5,3,8,6,4,11,9))+
  geom_errorbar(aes(ymin = asymp.LCL, ymax = asymp.UCL),
                width = 0.09, size  = 0.3)+
  ggtitle("Experiment I")+
  scale_y_continuous(
    breaks = seq(0, 70, 5),
    labels = seq(0, 70, 5))+
  geom_point(shape = 20,
             size  = 3) +
  theme_test(base_size =12, base_family = "serif")+
  theme(axis.title.y = element_text(margin = margin(t = 0,
  r = 5, b = 5, l = 0)))+
  theme(axis.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 90, vjust = .7,
        color = "black",size = 8),
        axis.text.y = element_text(color = "black"),
        panel.spacing = unit(0, "cm"))+
  ylab("Number of eggs")+ xlab("Isolates")


#grup
```

```
ggplot(medias,aes(x=reorder(isol,media), y=media))+
  geom_col(fill= c(8,4,5,5,2,8,2,5,2,8,
                   5,5,2,5,2,2,5,2,5,5,
                   5,2,5,8,2,4,8,4))+
  geom_errorbar(aes(ymin = asymp.LCL, ymax = asymp.UCL),
                width = 0.09, size  = 0.3)+
  ggtitle("Experiment I")+
  scale_y_continuous(
    breaks = seq(0, 70, 5),
    labels = seq(0, 70, 5))+
  geom_point(shape = 20,
             size  = 3) +
  theme_test(base_size =12, base_family = "serif")+
  theme(axis.title.y = element_text(margin = margin(t = 0,
  r = 5, b = 5, l = 0)))+
  theme(axis.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 90, vjust = .7,
        color = "black",size = 8),
        axis.text.y = element_text(color = "black")
        ,panel.spacing = unit(0, "cm"))+
  ylab("Number of eggs")+ xlab("Isolates")

#-------------------------------------------------------------------------
predito <- predict(model2, type = "response")
#residuo<-residuals(model2, type = "deviance")
AIC<-AIC(model2)
media<-tapply(predito,data$isol,mean)
var<-tapply(predito, data$isol, var)
sd<-tapply(predito, data$isol, sd)
med.ord<-sort(media)
media<-tapply(data$neggs,data$isol,mean)
med.ord<-sort(media)
bpp <- barplot(med.ord, beside=TRUE, border = "black",
               col = "lightgray",
               xlab = "isolates",
               ylab = "Deviance residuals  ",
               main = "Experiment I", las=2)
```

```
#Plotando os dados brutos e a curva esperada
#plot(predito~residuo, data=data, xlab="Deviance residuals",
ylab="predicted values")
#curve(exp(coef(model2)[1]+coef(model2)[2]*x),add=T)

bpp <- barplot(med.ord, beside=TRUE, border = "black",
                col = "lightgray",
                xlab = "isolates",
                ylab = "Deviance residuals  ",
                main = "Experiment I", las=2)

#-------------------------------------------------------------------------
library(ggplot2)
library(ggpubr)
library(factoextra)
data1<-data.frame(media,medias$SE)
n.cluster<-fviz_nbclust(scale(data1), kmeans,method = "wss")
n.cluster

# Compute k-means with k = 10
res.km <- kmeans(scale(data1), centers=10, nstart = 100)
print(res.km)

  # Visualize the clustering algorithm results.
km.clusters<-res.km$cluster
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))
#Creating the factor grouping 1
dados$grouping1 <- dados$isol
levels(dados$grouping1)
levels(dados$grouping1)<-c(2,10,8,8,1,4,1,8,7,9,5,8
,1,5,6,7,8,6,8,8,8,1,8,4,7,3,2,10)

#model fitted with factor grouping 1 instead of fung
model3 <- glm(neggs ~ block + grouping1 , family=quasipoisson,
data = dados)
summary(model3)
#Testing grouping (equality between isolates of the same group)
# LRT
```

```
anova(model3, model2, test = "F")

#get deviance for model
-2*logLik(model2)

#-------------------------------------------------------------------------------
# Compute k-means with k = 9
res.km <- kmeans(scale(data1), centers=9, nstart = 100)
print(res.km)

# Visualize the clustering algorithm results.
data.labels = data$isol
km.clusters<-res.km$cluster
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))

#Creating the factor grouping 2
dados$grouping2 <- dados$isol
levels(dados$grouping2)
levels(dados$grouping2)<-c(2,10,8,8,1,4,1,8,7,2,5,8,
1,5,6,7,8,6,8,8,8,1,8,4,7,3,2,10)

#model fitted with factor grouping 1 instead of fung
model4 <- glm(neggs ~ block + grouping2 , family=quasipoisson,
data = dados)

#Testing grouping (equality between isolates of the same group)
# LRT
anova(model4, model2, test = "F")
anova(model4, model3, test = "F")

#-------------------------------------------------------------------------------
# Compute k-means with k = 8
res.km <- kmeans(scale(data1), centers=8, nstart = 100)
print(res.km)

# Visualize the clustering algorithm results.
data.labels = data$isol
km.clusters<-res.km$cluster
```

```
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))

#Creating the factor grouping 3
dados$grouping3 <- dados$isol
levels(dados$grouping3)
levels(dados$grouping3)<-c(2,10,8,8,1,4,1,8,7,2,5,8,1,5,
1,7,8,1,8,8,8,1,8,4,7,3,2,10)

#model fitted with factor grouping 1 instead of fung
model5 <- glm(neggs ~ block + grouping3 , family=quasipoisson,
data = dados)

#Testing grouping (equality between isolates of the same group)
# LRT
anova(model5, model2, test = "F")
anova(model5, model3, test = "F")
anova(model5, model4, test = "F")
#-------------------------------------------------------------------------------
# Compute k-means with k = 7
res.km <- kmeans(scale(data1), centers=7, nstart = 100)
print(res.km)

# Visualize the clustering algorithm results.
data.labels = data$isol
km.clusters<-res.km$cluster
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))

#Creating the factor grouping 4
dados$grouping4 <- dados$isol
levels(dados$grouping4)
levels(dados$grouping4)<-c(2,10,8,8,1,4,1,8,7,2,8,
8,1,8,1,7,8,1,8,8,8,1,8,4,7,3,2,10)

#model fitted with factor grouping 1 instead of fung
model6 <- glm(neggs ~ block + grouping4 , family=quasipoisson,
data = dados)
```

```
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model6, model2, test = "F")
anova(model6, model3, test = "F")
anova(model6, model4, test = "F")
anova(model6, model5, test = "F")

#-----------------------------------------------------------------
# Compute k-means with k = 6
res.km <- kmeans(scale(data1), centers=6, nstart = 100)
print(res.km)

# Visualize the clustering algorithm results.
data.labels = data$isol
km.clusters<-res.km$cluster
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))

#Creating the factor grouping 5
dados$grouping5 <- dados$isol
levels(dados$grouping5)
levels(dados$grouping5)<-c(2,10,8,8,1,2,1,8,7,2,8,8,1,
8,1,7,8,1,8,8,8,1,8,2,7,3,2,10)

#model fitted with factor grouping 5 instead of fung
model7 <- glm(neggs ~ block + grouping5 , family=quasipoisson,
data = dados)

#Testing grouping (equality between isolates of the same group)
# LRT
anova(model7, model2, test = "F")
anova(model7, model3, test = "F")
anova(model7, model4, test = "F")
anova(model7, model5, test = "F")
anova(model7, model6, test = "F")

#-----------------------------------------------------------------
# Compute k-means with k = 5
res.km <- kmeans(scale(data1), centers=5, nstart = 100)
```

```
print(res.km)

# Visualize the clustering algorithm results.
data.labels = data$isol
km.clusters<-res.km$cluster
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))

#Creating the factor grouping 6
dados$grouping6 <- dados$isol
levels(dados$grouping6)
levels(dados$grouping6)<-c(2,10,8,8,1,2,1,8,1,2,8
,8,1,8,8,1,8,8,8,8,8,1,8,2,1,3,2,10)

#model fitted with factor grouping 5 instead of fung
model8 <- glm(neggs ~ block + grouping6 , family=quasipoisson,
data = dados)

#Testing grouping (equality between isolates of the same group)
# LRT
anova(model8, model2, test = "F")
anova(model8, model3, test = "F")
anova(model8, model4, test = "F")
anova(model8, model5, test = "F")
anova(model8, model6, test = "F")
anova(model8, model7, test = "F")

#-----------------------------------------------------------------
# Compute k-means with k = 4
res.km <- kmeans(scale(data1), centers=4, nstart = 100)
print(res.km)

# Visualize the clustering algorithm results.
data.labels = data$isol
km.clusters<-res.km$cluster
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))

#Creating the factor grouping 7
```

```
dados$grouping7 <- dados$isol
levels(dados$grouping7)
levels(dados$grouping7)<-c(2,10,8,8,1,2,1,8,1,2,8,8,1
,8,8,1,8,8,8,8,8,1,8,2,1,10,2,10)

#model fitted with factor grouping 5 instead of fung
model9 <- glm(neggs ~ block + grouping7 , family=quasipoisson,
data = dados)

#Testing grouping (equality between isolates of the same group)
# LRT
anova(model9, model2, test = "F")
anova(model9, model3, test = "F")
anova(model9, model4, test = "F")
anova(model9, model5, test = "F")
anova(model9, model6, test = "F")
anova(model9, model7, test = "F")
anova(model9, model8, test = "F")

#-------------------------------------------------------------------------
# Compute k-means with k = 3
res.km <- kmeans(scale(data1), centers=3, nstart = 100)
print(res.km)

# Visualize the clustering algorithm results.
data.labels = data$isol
km.clusters<-res.km$cluster
res.km$size
fviz_cluster(list(data=scale(data1), cluster = km.clusters))

#Creating the factor grouping 8
dados$grouping8 <- dados$isol
levels(dados$grouping8)
levels(dados$grouping8)<-c(2,2,8,8,1,8,1,8,1,2,8,
8,1,8,1,1,8,1,8,8,8,1,8,8,1,2,2,2)

#model fitted with factor grouping 5 instead of fung
model10 <- glm(neggs ~ block + grouping8 , family=quasipoisson,
data = dados)
```

```
#Testing grouping (equality between isolates of the same group)
# LRT
anova(model10, model2, test = "F")
anova(model10, model3, test = "F")
anova(model10, model4, test = "F")
anova(model10, model5, test = "F")
anova(model10, model6, test = "F")
anova(model10, model7, test = "F")
anova(model10, model8, test = "F")
anova(model10, model9, test = "F")


hnp(model8, print.on = T,pch= 4, main = "(A) Quasi-Poisson:
experiment I",cex=.5, cex.main=.9, pty="s", xlab = "Half-nomal scores",
ylab = "Deviance reriduals")
```

Apêndice II: computational routines

We carried out all programming in R For Poisson, Quasi-Poisson, negative binomial, Zero-inflated Poisson and Zero-inflated negative binimial models the following codes were used for number of flowers for experiment I:

```
source("helper01_general-functions.R")
source("helper02_lattice-panels.R")

# Predictor
f1 <- NF ~ block + trt + days
f2 <- NF ~ block + trt + days + I(days^2)
f3 <- NF ~ block + trt * (days + I(days^2))
f4 <- NF ~ block + trt*(days + I(days^2) + I(days^3))
#-------------------------------------------------------------------------------
# Poisson
modeloPO <- glm(f1, family = poisson, data = flower)
modelo1PO <- glm(f2, family = poisson, data = flower)
modelo2PO <- glm(f3, family = poisson, data =  flower)
modelo3PO <- glm(f4, family = poisson, data = flower)
sum(resid(modelo1PO, ty="pearson")^2)
sum(resid(modelo2PO, ty="pearson")^2)
```

```
summary(modelo2PO)
anova(modelo2PO, test = "Chisq")

par(mfrow=c(2,2))
hnp(modeloPO, paint.out = T, print.on = T, pch=4,
    main = "(a) Poisson; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo1PO, paint.out = T, print.on = T, pch=4,
    main = "(b) Poisson; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo2PO, paint.out = T, print.on = T, pch=4,
    main = "(a) Poisson; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo3PO, paint.out = T, print.on = T, pch=4,
    main = "(d) Poisson; experiment I",cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")

#-------------------------------------------------------------------------
# Quasi-Poisson
modeloQP <- glm(f1, family = quasipoisson, data = flower)
modelo1QP <- glm(f2, family = quasipoisson, data = flower)
modelo2QP <- glm(f3, family = quasipoisson, data = flower)
modelo3QP <- glm(f4, family = quasipoisson, data = flower)
#(phi <- 2442.468/ )

par(mfrow=c(2,2))
hnp(modeloQP, paint.out = T, print.on = T, pch=4,
    main="(a) Quasi-Poisson; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo1QP, paint.out = T, print.on = T, pch=4,
    main="(b) Quasi-Poisson; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo2QP, paint.out = T, print.on = T, pch=4,
    main="(a) Quasi-Poisson; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo3QP, paint.out = T, print.on = T, pch=4,
    main="(d) Quasi-Poisson; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
#-------------------------------------------------------------------------
```

```
# negative binomial
modeloBN <- glm.nb(f1, data = flower)
modelo1BN <- glm.nb(f2, data = flower)
modelo2BN <- glm.nb(f3, data = flower)
modelo3BN <- glm.nb(f4, data = flower)
thetahat <- summary(modelo2BN)$theta
thetahat
par(mfrow=c(2,2))
hnp(modeloBN, paint.out = T, print.on = T, pch=4,
    main="(a) Negative Binomial; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo1BN, paint.out = T, print.on = T, pch=4,
    main=" Negative Binomial; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo2BN, paint.out = T, print.on = T, pch=4,
    main=" (a) Negative Binomial; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo3BN, paint.out = T, print.on = T, pch=4,
    main=" Negative Binomial; experiment I", cex=0.5, cex.main=0.9,
    pty='s', xlab="Half-normal scores", ylab="Deviance residuals")

#-------------------------------------------------------------------------
## Zero-inflated Poisson

library(hnp)
require(pscl)
modeloPO.z1 <- zeroinfl( NF ~ block + trt + days| trt + days,
data = flower)
modeloPO.z2 <- zeroinfl( NF ~ block + trt + days + I(days^2)| trt
+ days + I(days^2), data = flower)
modeloPO.z3 <- zeroinfl(NF ~ block + trt * (days + I(days^2))|
trt * (days + I(days^2)), data = flower)
modeloPO.z4 <- zeroinfl(NF ~ block + trt*(days + I(days^2) +
I(days^3))| trt*(days + I(days^2) + I(days^3)), data = flower)
modeloPO.z5 <- zeroinfl(NF ~ block + trt+(days + I(days^2) +
I(days^3))|trt+(days + I(days^2) + I(days^3)), data = flower)


summary(modeloPO.z1)
```

```
summary(modeloPO.z3)

par(mfrow=c(2,2))
hnp(modeloPO.z1, paint.out = T, print.on = T, pch=4,
    main="(a) Zero-inflated-Poisson; experiment I", cex=0.5,
    cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloPO.z2, paint.out = T, print.on = T, pch=4,
    main=" Zero-inflated-Poisson; experiment I", cex=0.5,
    cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloPO.z3, paint.out = T, print.on = T, pch=4,
    main=" (a) Zero-inflated-Poisson; experiment I", cex=0.5,
    cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloPO.z4, paint.out = T, print.on = T, pch=4,
    main=" Zero-inflated-Poisson; experiment I", cex=0.5,
    cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloPO.z5, paint.out = T, print.on = T, pch=4,
    main=" Zero-inflated-Poisson; experiment II", cex=0.5,
    cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")

#-------------------------------------------------------------------------
modeloNB.z1 <- zeroinfl(NF ~ block + trt + days|  trt + days,
data = flowers, dist = "negbin")
modeloNB.z2<- zeroinfl(NF ~ block + trt + days + I(days^2)| trt +
days + I(days^2), data = flowers, dist = "negbin")
modeloNB.z3 <- zeroinfl(NF ~ block + trt * (days + I(days^2))| trt
+(days + I(days^2)), data = flowers, dist = "negbin")
modeloNB.z4 <- zeroinfl(NF ~ block + trt*(days + I(days^2) +
I(days^3))| trt*(days + I(days^2) + I(days^3)), data = flowers,
dist = "negbin")
modeloNB.z5 <- zeroinfl(NF ~ block + trt+ days + I(days^2) +
I(days^3)| trt + days + I(days^2) + I(days^3), data = flowers,
dist = "negbin")
summary(modeloNB.z5)
summary(modeloNB.z4)
```

```
par(mfrow=c(2,2))
hnp(modeloNB.z1, paint.out = T, print.on = T, pch=4,
    main="(a) Zero-inflated-Neg Binomial; experiment II",
    cex=0.5, cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloNB.z2, paint.out = T, print.on = T, pch=4,
    main="(b) Zero-inflated-Neg Binomial; experiment II",
    cex=0.5, cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloNB.z3, paint.out = T, print.on = T, pch=4,
    main="(a) Zero-inflated-Neg Binomial; experiment I",
    cex=0.5, cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloNB.z4, paint.out = T, print.on = T, pch=4,
    main="(d) Zero-inflated-Neg Binomial; experiment II",
    cex=0.5, cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")
hnp(modeloNB.z5, paint.out = T, print.on = T, pch=4,
    main="(e) Zero-inflated-Neg Binomial; experiment II",
    cex=0.5, cex.main=0.9, pty='s', xlab="Half-normal scores",
    ylab="Deviance residuals")

    #-------------------------------------------------------------------
predito <- predict(modeloNB.z3, type = "response")
media <- tapply(predito, flowers$trt, mean)
med.ord <- sort(media)
par(mar = c(6.5, 6.5, 2, 1) + 0.1)
bpp <- barplot(med.ord, beside=TRUE, border = "black",
               ylim = c(0,3.5),
               col = "lightgray",
               xlab = "isolates",
               ylab = "Predicted values of number of flowers  ",
               main = "Experiment I", las=2)


#-------------------------------------------------------------------
#Creating the factor grouping 1
```

```
flowers$grouping1 <- flowers$trt
levels(flowers$grouping1)
levels(flowers$grouping1)<-c(5,2,8,5,5,7,2,3,5,1,4,5,4,4,5,7,
2,5,6,3,6,2,3,3,4,4,5,3)


#model fitted with factor grouping 1 instead of fung
modeloNBG.z2 <- zeroinfl(NF ~ block + grouping1 + trt +
(days + I(days^2))| trt+(days+I(days^2)),
        dist = "negbin", data = flowers)

#-------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
library(lmtest)
lrtest( modeloNBG.z2, modeloNB.z2)
# The model5 is selected, because p>0.05 , the isolates within
#the groups do not differ statistically.
#-------------------------------------------------------------
#-------------------------------------------------------------
#Creating the factor grouping 2
flowers$grouping2 <- flowers$trt
levels(flowers$grouping2)
levels(flowers$grouping2)<-c(5,2,8,5,5,7,2,3,5,2,4,5,4,4,5,7,
2,5,6,3,6,2,3,3,4,4,5,3)

#model fitted with factor grouping 2 instead of fung
modeloNBG.z3 <- zeroinfl(NF ~ block + grouping2 + trt +
(days + I(days^2))| trt +(days + I(days^2)) ,
dist = "negbin", data = flowers)
#-------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
lrtest(modeloNBG.z3, modeloNBG.z2)

#-------------------------------------------------------------

#Creating the factor grouping 3
flowers$grouping3 <- flowers$trt
```

```
levels(flowers$grouping3)
levels(flowers$grouping3)<-c(5,2,8,5,5,7,2,4,5,2,4,5,4,4,5,7,2
,5,6,4,6,2,4,2,4,4,5,4)

modeloNBG.z4 <- zeroinfl(NF ~ block + grouping3 +
trt + (days + I(days^2))| trt +(days + I(days^2)) ,
dist = "negbin", data = flowers)
#
#-----------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
lrtest(modeloNBG.z4, modeloNBG.z3)



#-----------------------------------------------------------------
#Compute k-means with k = 5

#Creating the factor grouping 4
flowers$grouping4 <- flowers$trt
levels(flowers$grouping4)
levels(flowers$grouping4)<-c(5,2,8,5,5,7,2,4,5,2,4,5,4,4,5,7,
2,5,5,4,5,2,4,2,4,4,5,4)



#model fitted with factor grouping 4 instead of fung
modeloNBG.z5 <- zeroinfl(NF ~ block + grouping4 + trt +
(days + I(days^2))| trt +(days + I(days^2)) ,
dist = "negbin", data = flowers)

#-----------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
waldtest(modeloNBG.z6, modeloNBG.z3, test = "F")
lrtest(modeloNBG.z5, modeloNBG.z4)

#-----------------------------------------------------------------
#Creating the factor grouping 4
flowers$grouping5 <- flowers$trt
levels(flowers$grouping5)
```

```
levels(flowers$grouping5)<-c(5,2,8,5,5,7,2,4,5,2,4,5,4,4,5,7,2,
5,5,4,5,2,4,2,4,4,5,4)

#model fitted with factor grouping 4 instead of fung
modeloNBG.z6 <- zeroinfl(NF ~ block + grouping5 + trt +
(days + I(days^2))| trt +(days + I(days^2)) ,
dist = "negbin", data = flowers)


#-------------------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
lrtest(modeloNBG.z6, modeloNBG.z5)


#-------------------------------------------------------------------------
#Creating the factor grouping 4
flowers$grouping6 <- flowers$trt
levels(flowers$grouping6)
levels(flowers$grouping6)<-c(5,2,8,5,5,7,2,5,5,2,5,5,5,5,5,7,2,
5,5,5,5,2,5,2,5,5,5,5)


#model fitted with factor grouping 4 instead of fung
modeloNBG.z7 <- zeroinfl(NF ~ block + grouping6 + trt +
(days + I(days^2))| trt +(days + I(days^2)) ,
dist = "negbin", data = flowers)
#

#-------------------------------------------------------------------------
#Testing grouping (equality between isolates of the same group)
# LRT
lrtest(modeloNBG.z7, modeloNBG.z6)
```

Apêndice III: computational routines

We carried out all programming in R For Poisson, Quasi-Poisson, negative binomial, and COM-Poisson models the following codes were used for number of leaves for experiment I:

```
source("helper01_general-functions.R")
source("helper02_lattice-panels.R")

# Predictor
f1 <- NF ~ block + trt + days
f2 <- NF ~ block + trt + days + I(days^2)
f3 <- NF ~ block + trt * (days + I(days^2))
f4 <- NF ~ block + trt*(days + I(days^2) + I(days^3))

# Poisson
modeloPO <- glm(f1, family = poisson, data = leaves)
modelo1PO <- glm(f2, family = poisson, data = leaves)
modelo2PO <- glm(f3, family = poisson, data = leaves)
modelo3PO <- glm(f4, family = poisson, data = leaves)
anova(modeloPO, modelo1PO, test = "Chisq")
anova(modeloPO, test = "Chisq")
summary(modeloPO)

par(mfrow=c(2,2))
hnp(modeloPO, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I", cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo1PO, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I", cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo2PO, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I", cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo3PO, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I", cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")

# Quasi-Poisson
modeloQP <- glm(f1, family = quasipoisson, data = leaves)
modelo1QP <- glm(f2, family = quasipoisson, data = leaves)
modelo2QP <- glm(f3, family = quasipoisson, data = leaves)
modelo3QP <- glm(f4, family = quasipoisson, data = leaves)
anova(modeloQP, modelo1QP, test = "F")
anova(modeloQP, test = "F")
```

```
summary(modeloQP)

par(mfrow=c(2,2))
hnp(modeloQP, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I",cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo1QP,paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I",cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo2QP, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I",cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo3QP,paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I", cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")

# negative binomial
modeloBN <- glm.nb(f1, data = leaves)
modelo1BN <- glm.nb(f2, data = leaves)
modelo2BN <- glm.nb(f3, data = leaves)
modelo3BN <- glm.nb(f4, data = leaves)
getAnova(modeloBN, modelo1BN)
anova(modeloBN, modelo1BN)
anova(modeloBN, test = "F")
summary(modeloBN)

par(mfrow=c(2,2))
hnp(modeloBN, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I", cex=0.5,
cex.main=0.9, pty='s', xlab="Half-normal scores",
ylab="Deviance residuals")
hnp(modelo1BN, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I",cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo2BN, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I",cex=0.5, cex.main=0.9,
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
hnp(modelo3BN, paint.out = T, print.on = T, pch=4,
main="(a) Poisson; experiment I"cex=0.5, cex.main=0.9,
```

```
pty='s', xlab="Half-normal scores", ylab="Deviance residuals")
```