Joint modelling of mean and dispersion in dose-response experiments

Maíra B. Fatoretto [1], **Rafael A. Moral** [2], **Caroline Brophy** [2], **Clarice G. B. Demétrio** [1], **Joel C. Couceiro** [1]

Abstract: *One of the interests of fungus germination data is to assess the Ultraviolet radiation (UV-B) tolerance of reproductive structures (conidia) of entomopathogenic fungi. The ultraviolet radiation (UV-B) component reduces efficacy of fungi and the proportion of germinated conidia. The conidia germination is valued to select the most tolerant fungal isolate that could be used for the development of biopesticides. The proportion of germinated conidia in twelve isolates exposed to UV-B was observed aiming to identify the ones with highter resistance. Germination data are usually overdispersed. This may be due to individual variability of the experimental units, or due to some correlation between these units. We illustrate applications of Quasi-binomial and Beta-binomial models with the dispersion parameter modelled by covariates, using fungal germination data, in order to compare and to select the most suitable model to analyse and to compare the resistance of isolates. The Beta-binomial model with a dispersion parameters modelled using GAMLSS approach provided a good fit to the data and presented better results than Quasi-binomial and the Beta-binomial models.*

Keywords: germination; overdispersion; Quasi-binomial; Beta-binomial; GAMLSS.

[1]ESALQ, Universidade de São Paulo, Piracicaba, Brazil. e-mail: *mairafatoretto@gmail.com*

[2]Dept. of Mathematics and Statistics, Maynooth University, Maynooth, Ireland. e-mail: *rafael.deandradeMoral@mu.ie*

[2]Dept. of Mathematics and Statistics, Maynooth University, Maynooth, Ireland. e-mail: *caroline.brophy@mu.ie*

[1]ESALQ, Universidade de São Paulo, Piracicaba, Brazil. e-mail: *clarice.demetrio@usp.br*

[1]ESALQ, Universidade de São Paulo, Piracicaba, Brazil. e-mail: *joel.couceiro@usp.br*

Introduction

The class of generalized linear models (GLM) was introduced by Nelder and Wedderburn (1972) as a general framework for handling a range of statistical models for normal and non-normal data. The GLM framework allows us to analyze proportion data. An extension of these models are the Double generalized linear model which allow modelled simultaneously the mean and dispersion parameter in a generalized linear model. The efficacy of fungi is affected by ultraviolet radiation (UV-B). The incorporation of adjuvants in formulations (Isolates) can increase the protection to UV-B. One of the interests of fungus studies is to assess the UV-B tolerance of reproductive structures (conidia) of fungi using dose-response experiments, and these data are usually overdispersed. *Metarhizium spp.* is a diverse group of entomopathogenic fungi and this type of study improves the understanding of their adaptations to the environment and can assist in better selection of biopesticides. In this paper, we compare models to properly analyze overdispersed proportion data. The practical objective is to assess which isolate presented higher resistance to UV-B radiation.

Materials

To compare the efficacy of 12 Isolates of *Metarhizium spp.*, an experiment was conducted in a randomised complete block design at the 'Luiz de Queiroz' College of Agriculture (ESALQ), University of Sao Paulo (USP). In each of the blocks, one conidia suspension was prepared for each isolate and used for inoculation in 5 Petri dishes corresponding to each exposure time, adding up to 12 isolates $\times$ 5 exposure times (hours) $\times$ 3 blocks (blocks) = 180 observations. The response is the proportion of germinated conidia. At the beginning of the experiment, the isolate is less exposed to UV-B radiation, thus, has a higher germination rate. However, the longer the exposure time, greater is the damage to the isolate, which causes the germination reduction and, in some cases, the death of the conidia.

Methods

The class of generalized linear models (GLM) was introduced by Nelder and Wedderburn (1972) as a general framework for handling a range of common statistical models for normal and non-normal data. These models consist of three components:

(i) a univariate response variable with an associated distribution belonging to the exponential family (random component), with probability density function (pdf)

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

(ii) a linear predictor related to the explanatory variables

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $\boldsymbol{X} = [\boldsymbol{x_1},\ldots,\boldsymbol{x_n}]'$ is the $n \times p$ design matrix, and

(iii) a link function $\eta_i = g(\mu_i)$ relating the systematic to the random component. We have $\mu_i = \mathbb{E}[Y_i] = b'(\theta_i)$ and $\text{Var}[Y_i] = \phi b''(\theta_i) = \phi V(\mu_i)$, where $V(\mu_i)$ is called a variance function.

0.1 Quasi-binomial

The simplest way to deal with overdispersion is based in a quasi-likelihood approach, which requires the specification of the first two moments of the distribution (DEMETRIO; HINDE; MORAL, 2014). For the binomial GLM, $\text{Var}[Y_i]$ is replaced by

$$\text{Var}[Y_i] = \phi m_i \pi_i (1 - \pi_i), \tag{1}$$

where ϕ is called a dispersion parameter.

A usual way of estimating ϕ is through

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}. \tag{2}$$

0.2 Beta-binomial

Another way is considering the model Beta-binomial. Let $Y_i|P_i \sim \text{Binomial}(m_i, P_i)$ and $P_i \sim \text{Beta}(a_i, b_i)$, then marginally Y_i has a Beta-binomial distribution with mean and variance given by

$$\mathbb{E}[Y_i] = m_i \pi_i \quad \text{and} \quad \text{Var}[Y_i] = m_i \pi_i (1 - \pi_i)[1 + \phi(m_i - 1)]$$

where $\pi_i = \dfrac{a_i}{a_i + b_i}$, $\phi_i = \dfrac{1}{a_i + b_i + 1}$, with $a_i + b_i$ considered as constant. This model is often used when the probability of success is variable.

0.3 Beta-binomial using GAMLSS

Generalized Additive Models for Location, Scale and Shape (GAMLSS) were introduced by Rigby & Stasinopoulos (2001, 2005) and Akantziliotou et al. (2002) as a way of overcoming some limitations of the GLMs. One possility in this models if that the systematic part is expanded to allow modelling not only of the mean but other parameters of the distribution. Noteworthy GAMLSS is especially suited to modelling overdispersed data or which exhibit heterogeneity, where the scale or shape of the distributions changes with explanatory variables.

In the original parametrization the Beta-binomial in GAMLSS approach have the variance function given by

$$\text{Var}[Y_i] = m_i \pi_i (1 - \pi_i)[1 + \frac{\sigma}{1 + \sigma}(m_i - 1)] \tag{3}$$

0.4 Model fitting

We model overdispersion, verified by half-normal plots (MORAL; HINDE; DEMETRIO, 2014), and fitted a Quasi-binomial and Beta-binomial model (DEMETRIO; HINDE; MORAL, 2014). We also model the heterogeneity of variances between Isolates (Figure 1), through the modelling of the dispersion parameter, using Generalized Additive

Models for Location, Scale and Shape (GAMLSS) (RIGBY; STASINOPOULOS, 2005) in R software.

We began by fitting binomial, Quasi-binomial and Beta-binomial models with a 12 separated logistic regression lines. Due to the behavior of the data we also encompasses the quadratic term for dose, i.e.,

$$\eta_{ijk} = \beta_{0j} + \beta_{1j} t_k + \beta_{2j} t_k^2, \qquad \begin{matrix} i = 1, ..., 3 \\ j = 1, ..., 12 \\ k = 1, ..., 5 \end{matrix} \tag{4}$$

where β_{0j} and β_{1j} are the intercept and slope for the j-th isolate, respectively, and t_k is the k-th exposure time. The quadratic term was testing using the likelihood-ratio test and was significative, considering level of 5%.

Results

There is a difference between isolates and a large variability after four hours of exposure time in some of them (Figure 1). The Quasi-binomial model (M1) fitted the data well (Figure 2(a)), but it assumes that the dispersion parameter for all isolates is the same, which may not be realistic. We fitted a Beta-binomial model with the same and different dispersion per isolate (respectively, M2 and M3). We compared these models by the generalized Akaike criterion and log-likelihood. The likelihood ratio test (LRT) favours M3 (Table 1). Due to the similarity between some of the isolates, they can be grouped using LRT. The models presented different groupings, only M1 and M3 presented the same groupings, but diverged in some multiple comparisons. This should occur due to the difference in their confidence intervals (Figure 2(b)).

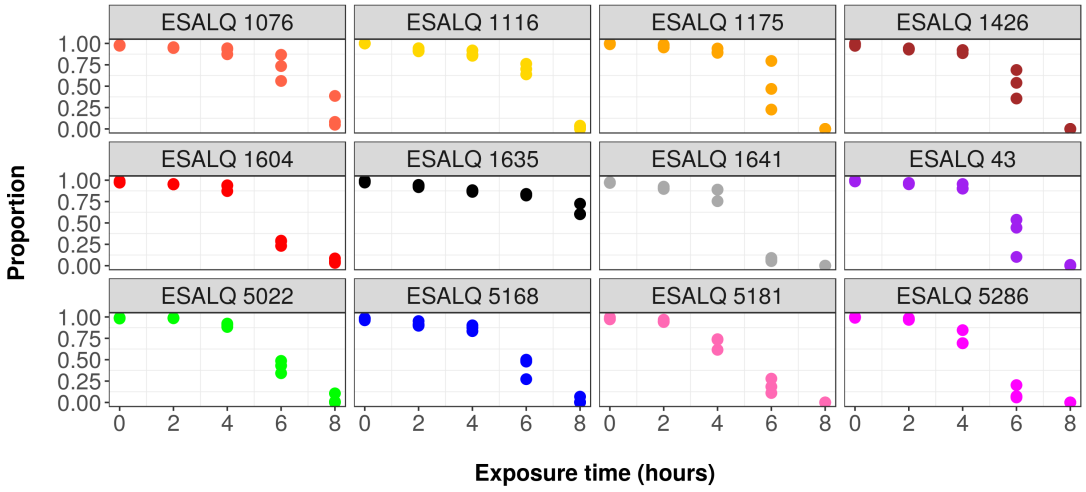

Figure 1: Plots of observed proportion vs. exposure time per isolate

Table 1: AIC and log-likelihood of Beta-binomial models.

Model	df	AIC	loglikelihood
M2	39	1322	-622.219
M3	40	1284	-602.219

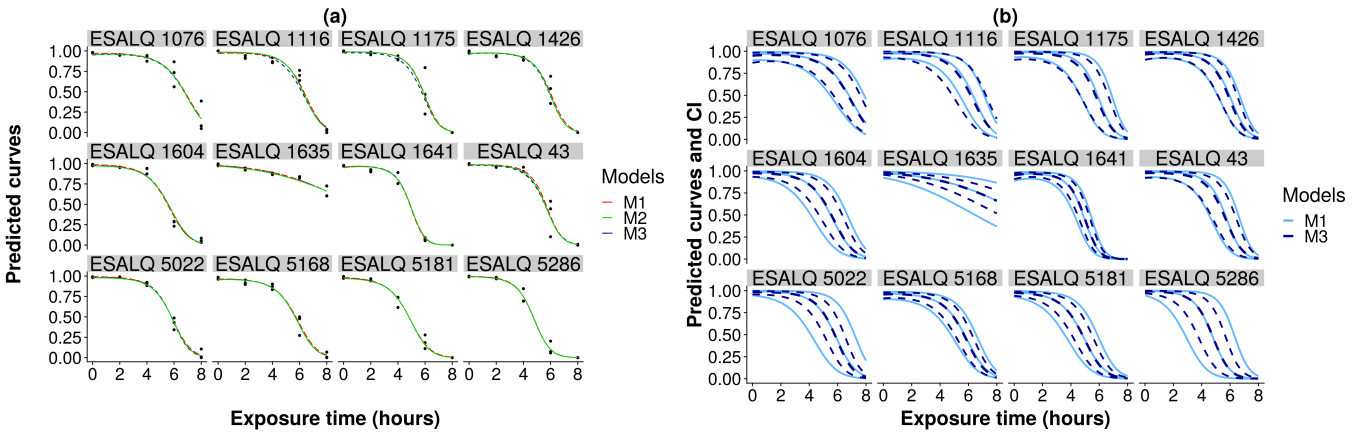

Figure 2: Fitted predicted curves using (a) (M1, M2 and M3) with observed data and (b) fitted predicted curves (M1 and M3) and respective confidence intervals.

Discussion

All models fitted well and presented similar predicted values, however model 3 presents more accurate confidence intervals discriminating the isolates better. As next steps, we will check these intervals using simulation and coverage rates (NEWCOMBE, 1998). We also will illustrate with other datasets of dose-response experiments how to improve the modelling of variance in proportion data.

References

AKANTZILIOTOU, K.; RIGBY, R.; STASINOPOULOS, D. The r implementation of generalized additive models for location, scale and shape. In: STATISTICAL MODELLING SOCIETY. *Statistical modelling in Society: Proceedings of the 17th International Workshop on statistical modelling*. [S.l.], 2002. p. 75–83.

DEMETRIO, C. G. B.; HINDE, J.; MORAL, R. de A. Models for overdispersed data in entomology. In: _____. First. [S.l.]: Springer, 2014. cap. 9, p. 219–259.

MORAL, R. de A.; HINDE, J.; DEMETRIO, C. G. B. *hnp: Half-Normal Plots with Simulation Envelopes*. [S.l.], 2014. R package version 1.0. Disponível em: <http://CRAN.R-project.org/package=hnp>.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, n. 3, p. 370–384, 1972.

NEWCOMBE, R. G. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in medicine*, Wiley Online Library, v. 17, n. 8, p. 873–890, 1998.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005.