Bioassays with natural mortality: handling overdispersion using random effects

Mariana Ragassi Urbano[1,2], *John Hinde* [1] *and Clarice Garcia Borges Demétrio* [2]

[1] *School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland, mrurbano@esalq.usp.br, john.hinde@nuigalway.ie*

[2] *Departamento de Ciências Exatas, ESALQ/USP, Piracicaba, São Paulo, Brazil, clarice@esalq.usp.br*

Abstract

In fitting dose-response models to entomological data it is often necessary to take account of natural mortality and/or overdispersion. The standard approach to handle natural mortality is to use Abbott's formula (Abbott, 1925), which allows for a constant underlying mortality rate. Standard overdispersion models include beta-binomial models, logistic-normal, and discrete mixtures. We extend the standard model (Morgan, 1992), and include a random effect in the dose levels, using the approach described in Aitkin et al. (2009). We consider the application of this model to data from an experiment on the use of a virus (PhopGV) for the biological control of worm larvae (Phthorimaea operculella) in potatoes, using a procedure implemented in software R. Using the model with random effects in the dose levels, we obtained a better fit than that provided by the standard model.

Introduction

Models for binary and binomial response grew out of the needs of a type of experimental investigation known as bioassay. In a typical bioassay, different concentrations of a chemical compound are applied to batches of experimental subjects and the number of subjects in each batch that respond to the chemical is then recorded. These values are regarded as observations on a binomial response variable. Some experiments in entomology exhibit evidence that responses can occur even at zero dose; here the response of interest is death and this phenomenon is referred to as *natural mortality*.

Among the available methodologies for the analysis of data that present natural mortality, little has been developed to deal with the occurrence of both natural mortality and overdispersion. To model situations like this, one approach is to use quasi-likelihood models, for example, as used by Raymond et al. (2006) and Mascarin et al. (2010).

In a bioassay, overdispersion can occur by variation in the response probabilities for groups of insects that received the same dose level. This variation could be attributed to relevant explanatory variables that have not been recorded, or to the inclusion in the model of certain variables that have not been adequately measured or controlled. According to Collet (2002), this situation can be modelled by the inclusion of a random effect in the linear predictor and so mixed models can be used in modelling overdispersion.

The application here is to an experiment in which potatoes (*Solanum tuberosum L.*) were each infected with $m_{ij} = 30$ larvae of *Phthorimaea operculella*, and then, D different concentrations of a virus (*PhopGV*) were applied to samples of n_i potatoes (observations are indexed by $i = 1, \ldots, D$ and $j = 1, \ldots, n_i$). There was also a control sample (no virus, $i = 0$) with $n_0 = 9$ potatoes. The experiment was conducted at 18^{o}C, and after 60 days the numbers of dead larvae y_{ij} were counted.

Methodology

In modelling the observed proportions y_{ij}/m_{ij}, the y_{ij} can be assumed to have a $B(m_{ij}, \pi^*_{ij})$ distribution, where π^*_{ij} the probability of response depends on the natural mortality and the dose-response relationship. A model for π^*_{ij} (Morgan,1992) is

$$\pi^*_{ij} = \omega_{ij} + (1-\omega_{ij})\pi_{ij}, \quad j = 1, ..., n_i \quad \text{and} \quad i = 0, ..., D,$$

where π_{ij} is given by the tolerance distribution (normal, logistic or extreme value), and ω_{ij} is the natural response probability. In general, we can model π_{ij} and ω_{ij} as function of covariates and parameters giving the following two models:

Model (a) - Standard model

$$\log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right) = \boldsymbol{\gamma'}\mathbf{u}_{ij} \quad \text{and} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta'}\mathbf{x}_{ij},$$

Model (b) - Random effect in the linear predictor of the dose levels

$$\log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right) = \boldsymbol{\gamma'}\mathbf{u}_{ij} \quad \text{and} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta'}\mathbf{x}_{ij} + \sigma Z_i,$$

where Z_i is a random effect with standard normal distribution.

If it were possible to label the subjects who responded due to the applied dose as y_{ijd} and those who responded naturally as y_{ijc}, then the total number of dead at dose d_{ij} would be

$$y_{ij} = y_{ijc} + y_{ijd}.$$

In the control group, if the number of larvae that died is y_{0j} out of m_{0j}, since they did not receive the virus we have $y_{0j} = y_{0jc}$.

The likelihood for Models (a) and (b) is given by

$$\begin{aligned} L(\omega_{ij}, \pi_{ij}; y_{ij}) &\propto \prod_{i=1}^{D}\prod_{j=1}^{n_i}[(1-\omega_{ij})(1-\pi_{ij})]^{m_{ij}-y_{ij}}\,[(1-\omega_{ij})\pi_{ij}]^{y_{ijd}}\omega_{ij}^{y_{ijc}} \\ &\times \prod_{j=1}^{n_0}\omega_{ij}^{y_{0j}}(1-\omega_{ij})^{m_{0j}-y_{oj}}. \end{aligned} \tag{1}$$

The log likelihood of (1) for Model (a) as function of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}) &\propto \sum_{i=1}^{D}\sum_{j=1}^{n_i}\left\{(m_{ij}-y_{ij})\left[\log\left(\frac{1}{1+e^{\boldsymbol{\beta'}\mathbf{x}_{ij}}}\right)\right] + y_{ijd}\log\left(\frac{e^{\boldsymbol{\beta'}\mathbf{x}_{ij}}}{1+e^{\boldsymbol{\beta'}\mathbf{x}_{ij}}}\right)\right. \\ &+ \left.(m_{ij}-y_{ij})\log\left(\frac{e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}{1+e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}\right) + y_{ijd}\log\left(\frac{1}{1+e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}\right) + y_{ijc}\log\left(\frac{e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}{1+e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}\right)\right\} \\ &+ \sum_{j=1}^{n_0} y_{0j}\log\left(\frac{e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}{1+e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}\right) + (m_{0j}-y_{0j})\log\left(\frac{1}{1+e^{\boldsymbol{\gamma'}\mathbf{u}_{ij}}}\right) \\ &= l(\boldsymbol{\beta}; \mathbf{y}) + l(\boldsymbol{\gamma}; \mathbf{y}). \end{aligned} \tag{2}$$

This log-likelihood is easy to maximize, because $l(\boldsymbol{\beta}; \mathbf{y}) + l(\boldsymbol{\gamma}; \mathbf{y})$ can be maximized separately. The approach used to estimate the parameters was the EM algorithm (Dempster et al., 1977), as also used

in bioassays with natural mortality by Hasselblad (1980) and Barlow and Feigl (1985). With the EM algorithm, the complete log-likelihood (2) is maximized iteratively by alternating between estimating y_{ijc} by its expectation under the current estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (E step) and then, with the y_{ijc}'s fixed at their expected values from the E step, maximizing $l(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y})$ (M-step). The $(k+1)^{th}$ iteration of the EM algorithm for Model (a) requires three steps:

$\mathbf{E-Step}$: Estimate $\mathrm{E}(y_{ijc}|y_{ij})$ under the current estimates $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$

$$y_{ijc}^{(k)} = \mathrm{E}(y_{ijc}|y_{ij}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}) = \left\{ \frac{e^{\boldsymbol{\gamma}'\mathbf{u}_{ij}} y_{ij}}{e^{\boldsymbol{\gamma}'\mathbf{u}_{ij}} + \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ij}}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_{ij}}}} \right. \quad \text{for } i = 1, ..., D;$$

$\mathbf{M-Step}$ for $\boldsymbol{\beta}$: Find $\boldsymbol{\beta}^{(k+1)}$ by maximizing $l(\boldsymbol{\beta}; y_{ijc}^{(k)}|y_{ij})$: $\boldsymbol{\beta}^{(k+1)}$ can be found from a binomial logistic regression of the responses $y_{ijd}^{(k)} = y_{ij} - y_{ijc}^{(k)}$ with binomial denominator $m_{ij} - y_{ijc}^{(k)}$ and design matrix $\mathbf{X}$;

$\mathbf{M-Step}$ for $\boldsymbol{\gamma}$: Find $\boldsymbol{\gamma}^{(k+1)}$ by maximizing $l(\boldsymbol{\gamma}; y_{ijc}^{(k)}|y_{ij})$: using a binomial logistic regression of the responses y_{0j} and $y_{ijc}^{(k)}$ with binomial denominators m_{0j} and m_{ij} respectively on design matrix $\mathbf{U}$.

These three steps must be repeated until the convergence is reached.
For Model (b), letting $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\sigma})$ be the combined parameter vector, the likelihood is given by

$$L(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=0}^{D} \left\{ \int_{-\infty}^{+\infty} \left[\prod_{j=1}^{n_i} P(y_{ij}|\boldsymbol{\psi}) \right] \phi(z_i) dz_i \right\}. \tag{3}$$

The integral in the likelihood (3) does not have a closed form except for normal Y. For other response models it is approximated by Gaussian quadrature: the integral over the normal Z_i is replaced by the finite sum over K Gaussian quadrature mass points z_k with masses α_k (Aitkin et al. 2009). The likelihood is then

$$L(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=0}^{D} \left\{ \sum_{k=1}^{K} \left[\prod_{j=1}^{n_i} P(y_{ij}|\boldsymbol{\psi}) \right] \alpha_k \right\},$$

where $P(y_{ij}|\boldsymbol{\psi}) = \binom{m_{ij}}{y_{ij}} (\pi_{ij}^*)^{y_{ij}} (1-\pi_{ij}^*)^{m_{ij}-y_{ij}}$.
The likelihood is thus (approximately) the likelihood of a finite mixture of exponential families density with known mixture proportions α_k at know mass-points z_k, thus z_k becomes another observable variable in the regression, with regression coefficient σ.
The log-likelihood is $l(\boldsymbol{\psi}; \mathbf{y}) = \sum_{i=0}^{D} \log \left(\sum_{k=1}^{K} \alpha_k \rho_{ik} \right)$, with $\rho_{ik} = \prod_{j=1}^{n_i} P(y_{ij}|\boldsymbol{\psi})$.
Then

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=0}^{D} \frac{\sum_{k=1}^{K} \alpha_k \rho_{ik} \frac{\partial \log \rho_{ik}}{\partial \boldsymbol{\beta}}}{\sum_{k=1}^{K} \alpha_k \rho_{ik}} = \sum_{i=0}^{D} \sum_{j=1}^{n_i} \sum_{k=1}^{K} w_{ik} \mathbf{s}_{ijk}(\boldsymbol{\beta}),$$

where w_{ik} is the posterior probability that observation y_{ij} comes from component k,

$$w_{ik} = \frac{\alpha_k \rho_{ik}}{\sum_{l=1}^{K} \alpha_k \rho_{il}}$$

and $\mathbf{s}_{ijk}(\boldsymbol{\beta})$ is the $\boldsymbol{\beta}$-component of the score function for observation (ij) in component k,

$$\mathbf{s}_{ijk}(\boldsymbol{\beta}) = \frac{(y_{ij} - \mu_{ijk})\mathbf{x}_{ij}}{\left(\frac{m_i \mu - \mu^2}{m_i}\right) g'_{ijk}}.$$

Following Anderson and Hinde (1988), the estimate of σ can be found by regarding Z_i as an additional covariate and σ as an extra parameter in the linear predictor. Estimation proceeds by fitting a weighted generalized linear model using w_{ik} as additional weights. These weights are functions of Z_i, Y_i, σ and $\boldsymbol{\beta}$ and must themselves be estimated iteratively.
The steps of the EM algorithm for model (b) are the following:

$\mathbf{E-Step}$: Estimate $\mathrm{E}(y_{ijc}|\, y_{ij})$ under the current estimates $\boldsymbol{\beta}^{(k)}$, $\sigma^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$,

$$y_{ijc}^{(k)} = \mathrm{E}(y_{ijc}|\, y_{ij}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}, \sigma^{(k)}) = \left\{ \frac{e^{\mathbf{u}_{ij}}\, y_{ij}}{e^{\mathbf{u}_{ij}} + \frac{e^{\boldsymbol{\beta}'\mathbf{x}_{ij}+\sigma z_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_{ij}+\sigma z_i}}} \right. \quad \text{for } i = 1, ..., D.$$

$\mathbf{M-Step}$ for $\boldsymbol{\beta}$ and σ: Find $\boldsymbol{\beta}^{(k+1)}$ and $\sigma^{(k+1)}$ by maximizing $l(\boldsymbol{\beta}, \sigma; y_{ijc}|y_{ij})$, $\boldsymbol{\beta}^{(k+1)}$ and $\sigma^{(k+1)}$ can be found from a weighted binomial regression of the responses $y_{ijd}^{(k)} = y_{ij} - y_{ijc}^{(k)}$ with binomial denominator $m_{ij} - y_{ijc}^{(k)}$ with weights w_{ik} for a design matrix $\mathbf{X}$ augmented by a vector $\mathbf{z}$ of the k Gaussian quadrature points.

$\mathbf{M-Step}$ for $\boldsymbol{\gamma}$: Find $\boldsymbol{\gamma}^{(k+1)}$ by maximizing $l(\boldsymbol{\gamma}; y_{ijc}|y_{ij})$: using a binomial logistic regression of the responses y_{0j} and $y_{ijc}^{(k)}$ with binomial denominators m_{0j} and m_{ij} respectively on design matrix $\mathbf{U}$.

In model (b) 20 quadrature points were used and the procedures were implemented in the R package.

Main Results and Conclusions

We included random effects in the standard model for natural mortality in the dose levels, with the aim to provide a better fit when the dataset exhibits overdispersion.
For the comparison between models, we also fitted a standard binomial model (Model (c)), with link functions logit and complementary log-log, without taking into account the natural mortality. Table 2 presents the fit statistics (-2 log likelihood, AIC and BIC) for models (a), (b) and (c).

Table 1: Fit Statistics: -2 log likelihood, AIC and BIC for models (a), (b) and (c)

Fit Statistics	Model (a) link function		Model (b) link function		Model (c) link function	
	logit	complementary log-log	logit	complementary log-log	logit	complementary log-log
-2 log likelihood	394.30	385.10	355.20	378.60	426.70	370.50
AIC	400.30	391.10	363.20	386.60	430.70	374.50
BIC	399.47	390.27	362.09	385.49	430.15	373.95

For these three statistics, the smaller the value the better is the fit. In Figure 1 we plot models (a), (b) and (c) with the link function logit, and in Figure 2 we plot models (a), (b) and (c) with the link function complementary log-log.
The fitted values for the models (b) were obtained by the empirical Bayes predictions (Aitkin, 1996).
The effective dose $ED_{100p\%}$ values, are doses which correspond under the model to $100p\%$ mortality. A commonly used summary of a fitted model is the ED_{50}, the dose corresponding to 50% mortality, that has a useful interpretation as the median of the tolerance distribution. In Table 2 are the values of the ED_{50} in log scale for the fitted models. For models (b), the approach used to calculate de ED_{50} is described in Gutreuter and Boogaard (2007).

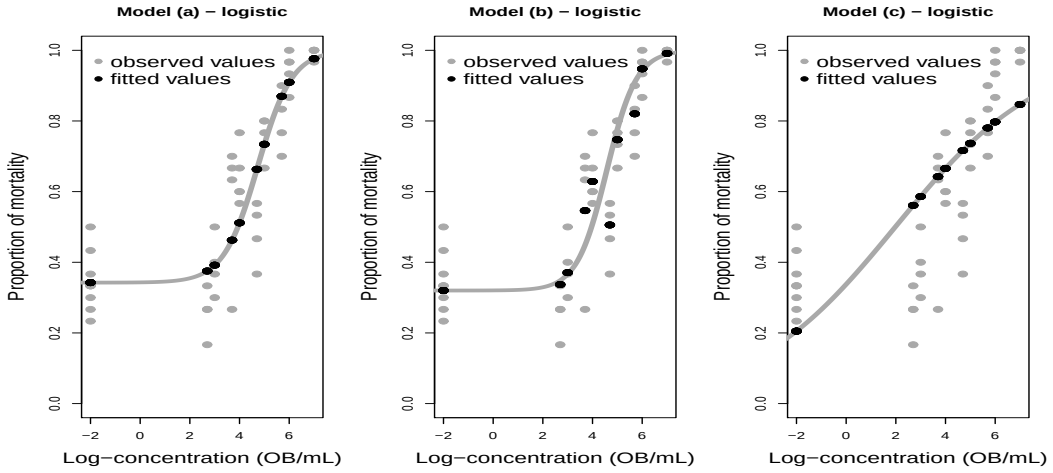

Figure 1: Proportion of mortality, fitted curve and predicted values for Models (a), (b) and (c) with logit link

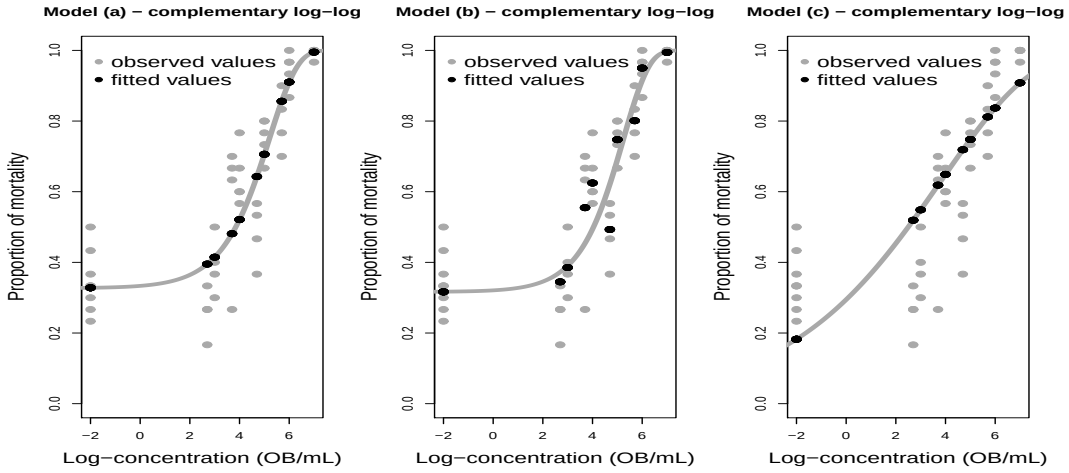

Figure 2: Proportion of mortality, fitted curve and predicted values for Models (a), (b) and (c) with complementary log-log link

Table 2: ED_{50} for models (a), (b) and (c)

	Model (a) link function		Model (b) link function		Model (c) link function	
	logit	complementary log-log	logit	complementary log-log	logit	complementary log-log
ED_{50}	4.73	4.80	4.57	4.86	1.97	2.50

With the inclusion of the random effect in the dose levels, the effective dose for the logistic model is 4.57 and for the complementary log-log model is 4.86. The ED_{50} when the natural mortality is not taken into account is underestimated, and when natural mortality is taken into account, but without random effect, the ED_{50} is overestimated, compared to the logistic model with random effect that provides the best fit. We concluded that data from biological assays that present natural mortality

and overdispersion can be more realistically modelled when a random effect is included to account for variability among the potatoes that received the same dose levels. For this dataset, the model with random effect in the linear predictor in the dose levels and logit link provides a better fit, and this model has response probability equation given by

$$\hat{\pi}_{ij} = 0.32 + (0.68)\frac{\exp[-7.61 + 1.66\log_{10}(d_{ij}) + 0.74z_i]}{1 + \exp[-7.61 + 1.66\log_{10}(d_{ij}) + 0.74z_i]}.$$

REFERENCES

Abbott, W.S. (1925). A method of computing the effectiveness of an insecticide. *Journal of Economic Entomology, 18, 265-267.*

Aitkin, M. (1996). Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. Statistical Modelling: *Proceedings of the 11th IWSM*.

Aitkin, M., Francis, B., Hinde, J. (2009). Statistical Modelling in R. *Oxford University Press, Oxford.*

Anderson, D.A., Hinde, J. (1988). Random effects in generalized linear models and the EM algorithm. *Communications in Statistics - Theory and Methods, 17, 3847-3856.*

Barlow, W.E., Feigl, P. (1985). Analysing binomial data with a nonzero baseline using GLIM. *Computational Statistics and Data Analysis, 3, 201-204.*

Collet, D. (2002). Modelling Binary Data. *Chapman and Hall/CRC.*

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society B, 39, 1-38.*

Gutreuter, S., Boogaard, A. (2007). Prediction of lethal/effective concentration/dose in the presence of multiple auxiliary covariates and components of variance. *Environmental Toxicology and Chemistry, 26, 1978-1986*.

Hasselblad, V., Stead, A.G., Creason, J.P. (1980). Multiple probit analysis with a nonzero background. *Biometrics, 36, 659-663*.

Mascarin, G.M., Alves, S.B., Ferreira, F.T.R., Urbano, M. R., Demétrio, C.G.B., Delalibera. I. (2010). Potential of a granulovirus isolate to control *Phthorimaea operculella* (Lepidoptera: Gelechiidae). *BioControl, 55, 657-671.*

Morgan, B.J.T. (1992). Analysis of Quantal Response Data. *Chapman and Hall/CRC.*

Raymond, B., Sayyed, A.H., Wright, D.J. (2006). The compatibility of a nucleopolyhedrosis virus control with resistance management for *Bacillus thuringiensis*: co-infection and cross-resistance studies with the diamondback moth, *Plutella xylostella*. *Journal of Invertebrate Pathology, 93, 114-120*

R (2011). Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0. URL `http://www.R-project.org`.

Acknowledgments: This work was supported by CAPES - Proc n^o 4942/10-8 and Science Foundation Ireland award 07/MI/012.