Brazilian Journal of Biometrics (2022), **40**, 490–509
doi:10.28951/bjb.v40i4.647

BRAZILIAN JOURNAL OF BIOMETRICS
ISSN:2764-5290

ARTICLE

Overdispersion Models for Clustered Toxicological Data in a Bioassay of Entomopathogenic Fungus

Silvia Maria de Freitas,[1] Lida Fallah,[2] Clarice G. B. Demétrio,[3] and John P. Hinde*,[4]

[1]Department of Statistics and Applied Mathematics, Federal University of Ceará¡, Fortaleza, CE, Brazil
[2]Discipline of Paediatrics, Trinity Translational Medicine Institute and School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3]Department of Exact Sciences, São Paulo University, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, SP, Brasil
[4]School of Mathematical and Statistical Sciences, University of Galway, Ireland
*Corresponding author. Email: john.hinde@universityofgalway.ie

(Received: June 1,2022; Revised: December 14,2022; Accepted: December 14,2022; Published: December 30,2022)

Abstract

We consider discrete mortality data for groups of individuals observed over time. The fitting of cumulative mortality curves as a function of time involves the longitudinal modelling of the multinomial response. Typically such data exhibit overdispersion, that is greater variation than predicted by the multinomial distribution. To model the extra-multinomial variation (overdispersion) we consider a Dirichlet-multinomial model, a random intercept model and a random intercept and slope model. We construct asymptotic and robust covariance matrix estimators for the regression parameter standard errors. Applying this model to a specific insect bioassay of the fungus *Beauveria bassiana*, we note some simple relationships in the results and explore why these are simply a consequence of the data structure. Fitted models are used to make inferences on the effectiveness and consistency of different isolates of the fungus to provide recommendations for its use as a biological control in the field.

Keywords: Grouped data; Dirichlet-multinomial; Extra-multinomial variation; Generalized estimating equations; Generalized linear models; Random effects models.

1. Introduction

Biological pest control uses entomopathogenic agents or pathogens (viruses, fungi, bacteria, parasites and nematodes) to control or eliminate a pest population. These biological control methods are an alternative to traditional methods and are becoming of increased interest because they provide ecologically and economically attractive replacements for chemical pesticides.

In toxicology experiments for these agents, it is common to use a response such as the cumulative mortality in a group (in a sampling context this is often referred to as a cluster) of insects measured at various time points (days) during the course of the experiment. Hence, the data is partly longitudinal since the cumulative mortality (for the individual groups) is modeled as a function of time and group level covariates. Petkau & Sitter (1989) drew attention to the problem of fitting cumulative mortality as a function of time to grouped data which involves modeling the multinomial response over time. They also emphasised an additional aspect, the possibility of extra-multinomial variation (overdispersion) arising from the use of groups of insects as the experimental units. A consequence of failing to take overdispersion into account is the underestimation of the standard errors of estimated regression coefficients giving incorrect significance of treatment effects (Hinde & Demétrio, 1998) and confidence intervals that are too narrow. In assessing the effectiveness of the different isolates of the pathogen we can consider the lethal time, *LTP*, until a certain percentage *P* of the insects have died; typically the *LT*50 (the median lethal dose) may be used as a simple summary of potency. Model-based estimates of these lethal times are easily obtained from the fitted regression coefficients and again, to obtain reliable estimates of the associated standard errors, we again need to take overdispersion into account. In selecting good isolates for consideration in field application we obviously require those with short lethal times, but the standard errors are also informative about the reliability of particular strains; as we will see some isolates exhibit much greater variability than others. In related studies with varying dose levels similar models can also help to determine application levels for use in the field.

Standard approaches for the analysis of this form of time course data include survival analysis (Petkau & Sitter, 1989) and ordinal generalized linear models (McCullagh, 1980; Glonek & McCullagh, 1995). However, these approaches are questionable when there is extra-multinomial variability in the data. The use of Liang & Zeger (1986) generalized estimating equations (GEE) as a multivariate version of quasi-likelihood (QL) (McCullagh & Nelder, 1989) has been considered as an approach to the problem of fitting a generalized linear model (GLM) to grouped/clustered data. The GEE approach addresses the problem of correlated or overdispersed data by using an adjusted score equation for the parameter estimates. We have to specify only the first two moments of the response vector for each group/cluster, assuming some form for the functional relationship between the mean and the variance. In GEE there is no need to specify the full likelihood, which is often intractable for non-Gaussian models even if additional assumptions are made. For Gaussian models the first and second moments fully identify the likelihood and so GEE is equivalent to maximum likelihood, but with non-Gaussian data the use of GEE provides estimating equations that are generally only optimal under specific assumptions about higher-order moments.

In this paper, we present four models suitable for grouped multinomial data, three of which are directed towards handling overdispersion. We use the GLM framework for multinomial data to model overdispersion in clustered data extending the approach given in O'Hara Hines & Lawless (1993), using a logit link function to relate the cumulative proportion to a specific linear predictor. The first model is the standard cumulative multinomial model. The second model is an extension to multinomial data of the beta-binomial model, which is often used for overdispersed binomial data. This model compounds the multinomial distribution for the observed counts with a Dirichlet distribution for the vector of underlying probabilities, leading to a Dirichlet-multinomial distribution, which has a different form for the variance function generalizing that for the standard multinomial distribution. Note that the Dirichlet-multinomial generalization has been applied in many different settings, including to nominal polytomous data (Salvador *et al.,* 2022), as an extension to deep learning for RNA sequencing data (Corsini & Viroli, 2022). The third model is a random intercept model, where we incorporate an additive random effect in the linear predictor to give a random location shift for each distinct multinomial sample. The final model is an extension of third model and includes both random intercepts and random slopes (Freitas, 2001).

The rest of the paper is organised as follows. In Section 2 we describe a data set from a biological control assay that we use to illustrate the proposed models. Details of the parameterizations and the variance-covariance structures of the models are discussed in Section 3. The procedures used to estimate the parameters are given in Section 4 and in Section 5 we analyse the data set and present the results. Section 6 discusses some particular aspects of the results that arise in applying the models to this specific regularly structured dataset. A theoretical derivation is provided to reinforce the validity of empirical results. The paper concludes with some additional comments and conclusions.

2. The biological dataset

The termite *Heterotermes tenuis* is an important pest of sugarcane in Brazil, causing damage of up to 10 metric tones/ha/year (Almeida *et al.,* 1997; Tamai, 1997), either during the planting season or the maturation phase. The fungus *Beauveria bassiana* is a possible microbial control for *H. tenuis* with the desirable characteristic of living inside the nests of the pest, which facilitates the dissemination of the pathogen. The data considered here are from a study conducted at the Insect Pathology Laboratory of ESALQ-USP, Piracicaba, São Paulo, Brazil. The aim was to determine the pathogenicity and virulence of 142 different isolates of *Beauveria bassiana*.

The data set discussed in this paper is from a completely randomized experiment, with five replicates of each of the 142 isolates. Solutions of the isolates (5×108 particles/ml) were applied to groups (clusters) of $n = 30$ termites kept in plastic Petri-dishes (60 mm diameter $\times$ 10 mm height). The mortality in the groups was measured daily for a period of eight days after application of the fungus, resulting in 710 multinomial observations of length eight. Figure 1 shows the cumulative proportion of dead termites against day for a sample of thirty isolates. Here we can see that there are clear differences in the efficacy of the isolates and evidence of differing degrees of variability between replicates.

As well as modelling the cumulative mortality as a function of time, a quantity of interest is the time until a certain percentage P of the termites have died, the lethal time LTP. However, it is also clear from Figure 1 that a single simple summary does not capture the full story and we may also be interested in selecting isolates that are highly potent, with short lethal times, and also reliably reproducible with low variation across replicates.

3. Models

Suppose that observations are taken on the isolates over D consecutive days and that initially there are n_{ik} insects for the k-th, $k = 1, \ldots, K$, replicate of isolate i, $i = 1, \ldots, I$. Then for isolate i, we write $Y_{ik,j}$ as the number of dead insects on day j ($j = 1, \ldots, D$) for the k-th replicate and $Y_{ik,D+1} = n_{ik} - \sum_{j=1}^{D} Y_{ik,j}$ as the number of insects still alive on day D. Let $R_{ik,j}$ denotes the cumulative proportion of insects dead by day j. Then $\boldsymbol{R}_{ik}$, the $D \times 1$ vector of cumulative proportions for replicate k of isolate i, is given by

$$\boldsymbol{R}_{ik} = (R_{ik,1}, R_{ik,2}, \ldots, R_{ik,\mathrm{D}})^T = \frac{1}{n_{ik}} \boldsymbol{L} \boldsymbol{Y}_{ik}, \tag{1}$$

where $\boldsymbol{L}$ is a $D \times (D + 1)$ matrix containing 1's on and below the diagonal and 0's above, and $\boldsymbol{Y}_{ik} = (Y_{ik,1}, Y_{ik,2}, \ldots, Y_{ik,\mathrm{D}+1})^T$. Note that $0 \leq R_{ik,1} \leq R_{ik,2} \leq \ldots \leq R_{ik,\mathrm{D}} \leq 1$.

3.1 Multinomial Model

In this paper we consider days as the multinomial categories with a natural increasing order from day $j = 1$ to $j = D$. The last category $D + 1$ includes all of the unobserved days, arising from

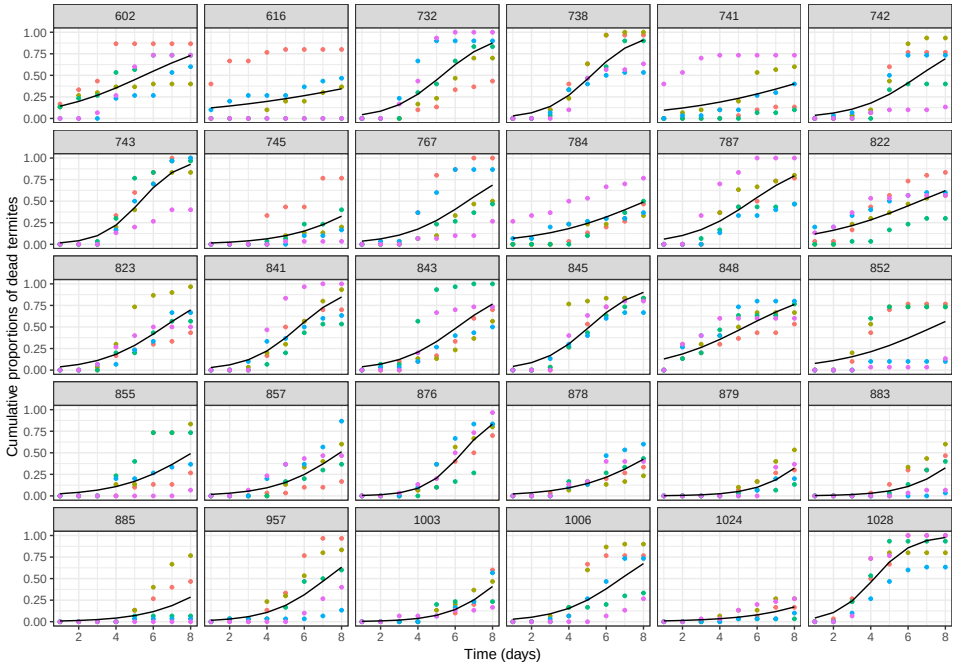

Figure 1. Examples of observed cumulative mortalilty proportions of *H. tenuis* over eight days for thirty typical isolates of *B. bassiana* showing the five replicates (in different colours) and fitted responses (black curves).

the censoring of observations at day D. Denoting the probability of an insect dying on day j for replicate k of isolate i by $\pi_{ik,j}$, and $\pi_{ik,D+1} = 1 - \sum_{j=1}^{D} \pi_{ik,j}$, we have $\boldsymbol{Y}_{ik} = (Y_{ik,1}, \ldots, Y_{ik,D}, Y_{ik,D+1})^T \sim$ Multinomial $(n_{ik}; \boldsymbol{\pi}_{ik})$, with the mean vector given by $\mathrm{E}[\boldsymbol{Y}_{ik}] = n_{ik}\boldsymbol{\pi}_{ik} = n_{ik}(\pi_{ik,1}, \ldots, \pi_{ik,D}, \pi_{ik,D+1})^T$ and covariance matrix

$$\mathrm{Var}[\boldsymbol{Y}_{ik}] = n_{ik}[\mathrm{diag}\{\boldsymbol{\pi}_{ik}\} - \boldsymbol{\pi}_{ik}\boldsymbol{\pi}_{ik}^T]. \tag{2}$$

In considering the cumulative proportions $R_{ik,j}$, since $R_{ij,D+1} \equiv 1$, we only need to use the first D of these derived random variables that give the information on the observed deaths, and so $R_{ij,D+1}$ is excluded from the subsequent modelling. From the moments of the multinomial distribution, using (1) and (2) we have

$$\mathrm{E}[\boldsymbol{R}_{ik}] = \mathrm{E}\left[\frac{1}{n_{ik}}\boldsymbol{L}\boldsymbol{Y}_{ik}\right] = \frac{1}{n_{ik}}\boldsymbol{L}\mathrm{E}[\boldsymbol{Y}_{ik}] = \boldsymbol{L}\boldsymbol{\pi}_{ik} = \boldsymbol{\gamma}_{ik},$$

where $\gamma_{ik,j}$ is the cumulative probability of an insect being dead by day j for replicate k of isolate i, and the covariance matrix is

$$\mathrm{Var}[\boldsymbol{R}_{ik}] = \mathrm{Var}\left[\frac{1}{n_{ik}}\boldsymbol{L}\boldsymbol{Y}_{ik}\right] = \frac{1}{n_{ik}}\boldsymbol{L}[\mathrm{diag}\{\boldsymbol{\pi}_{ik}\} - \boldsymbol{\pi}_{ik}\boldsymbol{\pi}_{ik}^T]\boldsymbol{L}^T = \boldsymbol{V}(\boldsymbol{\gamma}_{ik}), \tag{3}$$

where the matrix $\boldsymbol{V}(\boldsymbol{\gamma}_{ik})$ has elements $v_{jj'} = v_{j'j} = \gamma_{ik,j}(1 - \gamma_{ik,j'})/n_{ik}$, $(1 \le j \le j' \le D)$ (McCullagh & Nelder, 1989).

For a particular isolate i and replicate k, the vector of cumulative proportions $\boldsymbol{R}_{ik}$ will be used to model the vector of cumulative probabilities $\boldsymbol{\gamma}_{ik} = (\gamma_{ik,1}, \ldots, \gamma_{ik,D})^T$. Using a GLM for $\boldsymbol{\gamma}_{ik}$ we have $g(\boldsymbol{\gamma}_{ik}) = \boldsymbol{X}_{ik}\boldsymbol{\beta}_i$ for some suitable link function g (logit, probit or complementary log-log), where $\boldsymbol{\beta}_i$

is a $m \times 1$ vector of parameters to be estimated and $\boldsymbol{X}_{ik}$ is a $D \times m$ matrix of covariates. Here we use the logit link function which gives the cumulative logistic model

$$g(\gamma_{ik,j}) = \text{logit}(\gamma_{ik,j}) = \log\left(\frac{\gamma_{ik,j}}{1-\gamma_{ik,j}}\right) = \log\left(\frac{\sum_{s=1}^{j}\pi_{ik,s}}{\sum_{s=j+1}^{D+1}\pi_{ik,s}}\right) = \eta_{ik,j}.$$

The linear predictor may contain isolate specific factors and covariates to model the time dependency. The model used here is an isolate specific linear time effect

$$\eta_{ik,j} = \beta_{1i} + \beta_{2i}(t_j - \bar{t}),$$

where β_{1i} is the baseline effect for the i-th isolate (corresponding to initial mortality on day 0), $t_j = j$ is a quantitative variable for day j, $\bar{t}$ is the mean time, and the coefficient β_{2i} is the isolate specific time effect. Note that these coefficients are assumed to be constant over replicates. The lethal time for a death percentage P for the i-th isolate is

$$LTP_i = \frac{\text{logit}(P/100) - \beta_{1i}}{\beta_{2i}} + \bar{t}.$$

The mean-centered time version of the model is used for computational reasons and to allow the use of uncorrelated random effects in the model with random slope and random intercept. In the following, for simplicity of exposition, we will use t to denote the mean-centered time $(t - \bar{t})$.

3.2 Dirichlet-Multinomial Model (DM)

The standard multinomial distribution assumes that the n_{ik} responses are independent realisations with the same probability structure, e.g., probability $\pi_{ik,j}$ of dying on day j. However, because of the experimental setup we might expect correlation between the insects with the experimental units (the replicate petri dishes) or possibly variation across replicates because of how the treatments are applied. Both of these aspects will lead to additional variability across the replicates giving overdispersion relative to the standard variance function in (3). One way to allow for overdispersion is to adopt a two-stage model in which the multinomial parameter vector is assumed to have some distribution. This extension obviously directly allows for variation across replicates but can also be shown to induce correlation within clusters because of the shared random effects.

Here, as a first stage, we take as the response model, $\boldsymbol{Y}_{ik} \mid \boldsymbol{p}_{ik} \sim \text{Multinomial}(n_{ik}; \boldsymbol{p}_{ik})$, and then extend this to a second stage by allowing the cell probability vector $\boldsymbol{p}_{ik} = (p_{ik,1}, \ldots, p_{ik,\text{D}}, p_{ik,\text{D+1}})^T$ to follow a Dirichlet distribution (Mosimann, 1962) with probability density function given by

$$f(\boldsymbol{p}_{ik}; \boldsymbol{\alpha}_i) = \frac{\Gamma(\sum_{j=1}^{\text{D+1}}\alpha_{ij})}{\prod_{j=1}^{\text{D+1}}\Gamma(\alpha_{ij})} p_{ik,1}^{\alpha_{i1}-1} p_{ik,2}^{\alpha_{i2}-1} \ldots p_{ik,\text{D}}^{\alpha_{i\text{D}}-1} p_{ik,\text{D+1}}^{\alpha_{i\text{D+1}}-1}, \tag{4}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{i\text{D}}, \alpha_{i\text{D+1}})^T$, $0 < p_{ik,j} < 1$, $\alpha_{ij} > 0$, $j = 1, 2, \ldots, D+1$, $\sum_{j=1}^{\text{D+1}} p_{ik,j} = 1$,

$$\text{E}[\boldsymbol{p}_{ik}] = \frac{1}{(\sum_{j=1}^{\text{D+1}}\alpha_{ij})}\boldsymbol{\alpha}_i = \boldsymbol{\pi}_{ik} \tag{5}$$

and

$$\text{Var}[\boldsymbol{p}_{ik}] = [\text{diag}\{\boldsymbol{\pi}_{ik}\} - \boldsymbol{\pi}_{ik}\boldsymbol{\pi}_{ik}^T]\rho_i, \tag{6}$$

where $\rho_i = 1/(1 + \sum_{j=1}^{D+1} \alpha_{ij})$.

Unconditionally, $\boldsymbol{Y}_{ik}$ follows a Dirichlet-multinomial distribution (Mosimann, 1962) with probability density function given by

$$P(\boldsymbol{Y}_{ik}; n_{ik}) = \binom{n_i}{y_{ik,1}, \ldots, y_{ik,D+1}} \frac{\Gamma\left(\sum_{j=1}^{D+1} \alpha_{ij}\right)}{\Gamma\left(n_{ik} + \sum_{j=1}^{D+1} \alpha_{ij}\right)} \prod_{j=1}^{D+1} \frac{\Gamma\left(y_{ik,j} + \alpha_{ij}\right)}{\Gamma\left(\alpha_{ij}\right)} \tag{7}$$

where $0 \leq y_{ik,j} \leq n_{ik}$, $\alpha_{ij} > 0$.

Using standard results on conditional expectations and expressions (5) and (6) from the Dirichlet distribution, the mean vector and covariance matrix of the Dirichlet-multinomial model are given by

$$\text{E}[\boldsymbol{Y}_{ik}] = \text{E}[\text{E}(\boldsymbol{Y}_{ik}|\boldsymbol{p}_{ik})] = \frac{n_{ik}}{(\sum_{j=1}^{D+1} \alpha_{ij})}\boldsymbol{\alpha}_i = n_{ik}\boldsymbol{\pi}_{ik},$$

and

$$\text{Var}[\boldsymbol{Y}_{ik}] = n_{ik}\left[\text{diag}\{\boldsymbol{\pi}_{ik}\} - \boldsymbol{\pi}_{ik}\boldsymbol{\pi}_{ik}^T\right][1 + \rho_i(n_{ik} - 1)], \tag{8}$$

where $\pi_{ik,j} = \alpha_{ij}/(\sum_{j=1}^{D+1} \alpha_{ij})$ and ρ_i is an overdispersion parameter, $-1/(n_{ik} - 1) \leq \rho_i \leq 1$. Here, $[1 + \rho_i(n_{ik} - 1)]$ inflates the variance function of the multinomial distribution and when $n_{ik} = n$ for all i and k, $\rho_i = \rho$ it is of the same form as the heterogeneity factor of Finney (1971). Note that the variance function given by (8) remains valid when ρ_i is negative, provided that $\rho_i \geq -1/(n_{ik} - 1)$ for all k, although this requires some of the α's to be negative and hence can no longer be interpreted as a mixture distribution. However, additional constraints may also be required to ensure that the marginal probabilities given by (7) are non-negative.

In terms of the cumulative proportions vector $\boldsymbol{R}_{ik}$, given in (1), we have

$$\text{E}[\boldsymbol{R}_{ik}] = \text{E}\left[\frac{1}{n_{ik}}\boldsymbol{L}\boldsymbol{Y}_{ik}\right] = \frac{1}{n_{ik}}\boldsymbol{L}\text{E}[\text{E}(\boldsymbol{Y}_{ik}|\,\boldsymbol{p}_{ik})] = \boldsymbol{L}\boldsymbol{\pi}_{ik} = \boldsymbol{\gamma}_{ik},$$

and covariance matrix given by

$$\begin{aligned}
\text{Var}[\boldsymbol{R}_{ik}] &= \text{Var}\left[\frac{1}{n_{ik}}\boldsymbol{L}\boldsymbol{Y}_{ik}\right] = \frac{1}{n_{ik}^2}\boldsymbol{L}\{\text{E}[\text{Var}(\boldsymbol{Y}_{ik}|\boldsymbol{p}_{ik})] + \text{Var}[\text{E}(\boldsymbol{Y}_{ik}|\boldsymbol{p}_{ik})]\}\boldsymbol{L}^T \\
&= \frac{1}{n_{ik}}\boldsymbol{L}\left[\text{diag}\{\boldsymbol{\pi}_{ik}\} - \boldsymbol{\pi}_{ik}\boldsymbol{\pi}_{ik}^T\right]\boldsymbol{L}^T[1 + \rho_i(n_{ik} - 1)] \\
&= \boldsymbol{V}(\boldsymbol{\gamma}_{ik})[1 + \rho_i(n_{ik} - 1)],
\end{aligned} \tag{9}$$

where $\boldsymbol{V}(\boldsymbol{\gamma}_{ik})$ is given by (3). In the absence of overdispersion, e.g., $\rho_i = 0$, the covariance matrix (9) reduces to the multinomial cumulative structure given by (3).

Note that when $D = 1$, (4) is a Beta$(\alpha_{i1}, \alpha_{i2})$ distribution with $p_{ik,2} = 1 - p_{ik,1}$, $\text{E}(p_{ik,1}) = \alpha_{i1}/(\alpha_{i1} + \alpha_{i2})$ and $\text{Var}(p_{ik,1}) = \alpha_{i1}\alpha_{i2}/[(\alpha_{i1} + \alpha_{i2})^2(1 + \alpha_{i1} + \alpha_{i2})]$, and then (7) gives a beta-binomial model, with $\text{Var}(Y_{ik,j}) = n_{ik}\pi_{ik,j}(1 - \pi_{ik,j})[1 + (n_{ik} - 1)\rho_i]$, as described by Williams (1982) and Crowder (1978) for

overdispersed binomial data. Note that (9) also reduces to a beta-binomial model with $\text{Var}(R_{ik,j}) = \gamma_{ik,j}(1-\gamma_{ik,j})[1+(n_{ik}-1)\rho_i]/n_{ik}$.

In the following we consider using both a Dirichlet-multinomial model where the overdispersion, ρ_i, is allowed to vary over the isolates and also one with a single constant overdispersion term common to all isolates, that is with $\rho_i = \rho$, for all i.

3.3 Random Intercept Model (REM I)

We can see from Figure 1 that there is considerable variability among the replicates for some isolates, with changing level and slope for the cumulative proportion profiles. Indeed this is true for many of the full set of 142 isolates. We can attempt to allow for this in a model by assuming that the linear predictor, $\boldsymbol{\eta}_{ik}$, has some additional component variability. For example, incorporating an additive random effect into the linear predictor, gives a random location shift in the baseline of each isolate for each replicate and we have

$$g(q_{ik,j}) = \eta_{ik,j} + \xi_{ik} = \beta_{1i} + \beta_{2i}t_j + \xi_{ik}, \tag{10}$$

where $q_{ik,j} = \sum_{s=1}^{j} p_{ik,s}$ is the cumulative probability, ξ_{ik} is a random effect with $\text{E}[\xi_{ik}] = 0$, $\text{Var}[\xi_{ik}] = \sigma_i^2$ and the ξ_{ik}'s are independent. An obvious approach would be to assume a normal distribution for ξ_{ik}. Writing $q_{ik,j} = g^{-1}(\eta_{ik,j} + \xi_{ik}) = h(\eta_{ik,j} + \xi_{ik})$ and using a first-order Taylor series expansion of $h(\eta_{ik,j} + \xi_{ik})$ around the linear predictor $\eta_{ik,j}$, we obtain $q_{ik,j} \approx h(\eta_{ik,j}) + h'(\eta_{ik,j})\xi_{ik}$, where $h'(\eta_{ik,j}) = \partial h(\eta_{ik,j})/\partial\eta_{ik,j}$. Then $\text{E}[\boldsymbol{q}_{ik}] \approx \mathbf{h}(\boldsymbol{\eta}_{ik}) = \boldsymbol{\gamma}_{ik}$ and

$$\text{Var}[\boldsymbol{q}_{ik}] \approx [\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T\sigma_i^2,$$

where $\mathbf{h}'(\boldsymbol{\eta}_{ik}) = (h'(\eta_{ik,1}),\ldots,h'(\eta_{ik,D}))^T$ and $\boldsymbol{\gamma}_{ik} = (\gamma_{ik,1},\ldots,\gamma_{ik,D})^T$, showing that $\text{Var}(q_{ik,j}) \approx \sigma_i^2[h'(\eta_{ik,j})]^2$ and $\text{Cov}(q_{ik,j}, q_{ik,j'}) \approx \sigma_i^2 h'(\eta_{ik,j})h'(\eta_{ik,j'})$.

Then, the resulting marginal distribution for the cumulative response vector has mean vector

$$\text{E}[\boldsymbol{R}_{ik}] = \text{E}[\text{E}(\boldsymbol{R}_{ik}|\boldsymbol{q}_{ik})] = \text{E}[\boldsymbol{q}_{ik}] \approx \boldsymbol{\gamma}_{ik}$$

and covariance matrix given by

$$\begin{aligned}
\text{Var}[\boldsymbol{R}_{ik}] &= \text{Var}[\text{E}(\boldsymbol{R}_{ik}|\boldsymbol{q}_{ik})] + \text{E}[\text{Var}(\boldsymbol{R}_{ik}|\boldsymbol{q}_{ik})] \\
&\approx [\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T\sigma_i^2 + \text{E}[V(\boldsymbol{q}_{ik})] \\
&= V(\boldsymbol{\gamma}_{ik}) + \left(1 - \frac{1}{n_{ik}}\right)\sigma_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T,
\end{aligned} \tag{11}$$

since

$$\begin{aligned}
\text{E}[V(\boldsymbol{q}_{ik})] &= \frac{1}{n_{ik}}\text{E}\begin{pmatrix} q_{ik,1}-q_{ik,1}^2 & q_{ik,1}-q_{ik,1}q_{ik,2} & \cdots & q_{ik,1}-q_{ik,1}q_{ik,D} \\ q_{ik,1}-q_{ik,1}q_{ik,2} & q_{ik,2}-q_{ik,2}^2 & \cdots & q_{ik,2}-q_{ik,2}q_{ik,D} \\ \cdots & \cdots & \cdots & \cdots \\ q_{ik,1}-q_{ik,1}q_{ik,D} & q_{ik,2}-q_{ik,2}q_{ik,D} & \cdots & q_{ik,D}-q_{ik,D}^2 \end{pmatrix} \\
&= V(\boldsymbol{\gamma}_{ik}) - \frac{1}{n_{ik}}\sigma_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T,
\end{aligned}$$

where $\boldsymbol{V}(\boldsymbol{\gamma}_{ik})$ is given by (3). In the absence of overdispersion, with $\sigma_i^2 = 0$, the covariance matrix (11) reduces to the multinomial cumulative structure given by (3). When $D = 1$, (11) reduces to the approximate variance function for the logistic-normal model given by Williams (1982).

3.4 Random Intercept and Random Slope Model (REM II)

The model presented in (10) incorporates a random effect for the intercept. Another possibility is to extend this to include an additive random effect for the slope, giving a random slope effect between the replicates for each isolate (Freitas, 2001). Then, for each isolate we can write the linear predictor as

$$g(q_{ik,j}) = \beta_{1i} + \xi_{ik} + (\beta_{2i} + \zeta_{ik})t_j = \eta_{ik,j} + \xi_{ik} + \zeta_{ik}t_j,$$

where $(\xi_{ik}, \zeta_{ik})^T$ has a distribution with $\mathrm{E}[\xi_{ik}] = \mathrm{E}[\zeta_{ik}] = 0$ and covariance matrix

$$\boldsymbol{\Sigma}_{\nu_i\tau_i\lambda_i} = \begin{bmatrix} \nu_i^2 & \lambda_i\nu_i\tau_i \\ \lambda_i\nu_i\tau_i & \tau_i^2 \end{bmatrix},$$

where $-1 \leq \lambda_i \leq 1$ is the correlation between ξ_{ik} and ζ_{ik}. A particular case would be to assume a bivariate normal distribution for $(\xi_{ik}, \zeta_{ik})^T$. Writing

$$q_{ik,j} = g^{-1}(\eta_{ik,j} + \xi_{ik} + \zeta_{ik}t_j) = h(\eta_{ik,j} + \xi_{ik} + \zeta_{ik}t_j)$$

and using a first-order Taylor series expansion for the cumulative multinomial probability $q_{ik,j} = h(\eta_{ik,j} + \xi_{ik} + \zeta_{ik}t_j)$, around the linear predictor $\eta_{ik,j}$, we obtain $q_{ik,j} \approx h(\eta_{ik,j}) + h'(\eta_{ik,j})(\xi_{ik} + \zeta_{ik}t_j)$. Then $\mathrm{E}[\boldsymbol{q}_{ik}] \approx \mathbf{h}(\boldsymbol{\eta}_{ik}) = \boldsymbol{\gamma}_{ik}$ and

$$\begin{aligned} \mathrm{Var}[\boldsymbol{q}_{ik}] \approx\ & \nu_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T + \\ & \tau_i^2\{\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}\}\{\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}\}^T + \\ & \lambda_i\nu_i\tau_i[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T \odot [\mathbf{1}\mathbf{t}_{ik}^T + \mathbf{t}_{ik}\mathbf{1}^T], \end{aligned}$$

where $\mathbf{1}$ is a $D \times 1$ unit column vector, $\mathbf{t}_{ik}^T = (t_{ik,1}, t_{ik,2}, \ldots, t_{ik,D})$ and $\odot$ indicates the element by element product operation (Hadamard product, see for example Rao, 1973). So now we have $\mathrm{Var}[q_{ik,j}] \approx [h'(\eta_{ik,j})]^2[\nu_i^2 + \tau_i^2 t_j^2 + 2\lambda_i\nu_i\tau_i t_j]$ and $\mathrm{Cov}[q_{ik,j}, q_{ik,j'}] \approx h'(\eta_{ik,j})h'(\eta_{ik,j'})[\nu_i^2 + \tau_i^2 t_j t_{j'} + \lambda_i\nu_i\tau_i(t_j + t_{j'})]$.

The resulting marginal distribution for the cumulative response vector has mean vector

$$\mathrm{E}[\boldsymbol{R}_{ik}] = \mathrm{E}[\mathrm{E}(\boldsymbol{R}_{ik}|\boldsymbol{q}_{ik})] = \mathrm{E}[\boldsymbol{q}_{ik}] \approx \boldsymbol{\gamma}_{ik}$$

and covariance matrix given by

$$\begin{aligned} \mathrm{Var}[\boldsymbol{R}_{ik}] &= \mathrm{Var}[\mathrm{E}(\boldsymbol{R}_{ik}|\boldsymbol{q}_{ik})] + \mathrm{E}[\mathrm{Var}(\boldsymbol{R}_{ik}|\boldsymbol{q}_{ik})] \\ &\approx \nu_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T + \tau_i^2\{\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}\}\{\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}\}^T + \\ &\quad \lambda_i\nu_i\tau_i[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T \odot [\mathbf{1}\mathbf{t}_{ik}^T + \mathbf{t}_{ik}\mathbf{1}^T] + \mathrm{E}[\boldsymbol{V}(\boldsymbol{q}_{ik})] \\ &= \boldsymbol{V}(\boldsymbol{\gamma}_{ik}) + \left(1 - \frac{1}{n_{ik}}\right)\Big\{\nu_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T + \\ &\quad \tau_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}][\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}]^T + \\ &\quad \lambda_i\nu_i\tau_i[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T \odot [\mathbf{1}\mathbf{t}_{ik}^T + \mathbf{t}_{ik}\mathbf{1}^T]\Big\}, \end{aligned} \tag{12}$$

since

$$\begin{aligned}\mathrm{E}[V(\boldsymbol{q}_{ik})] &\approx \boldsymbol{V}(\boldsymbol{\gamma}_{ik}) - \frac{1}{n_{ik}}\nu_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T - \frac{1}{n_{ik}}\tau_i^2[\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}][\mathbf{h}'(\boldsymbol{\eta}_{ik}) \odot \mathbf{t}_{ik}]^T \\ &\quad -\frac{1}{n_{ik}}\lambda_i\nu_i\tau_i[\mathbf{h}'(\boldsymbol{\eta}_{ik})][\mathbf{h}'(\boldsymbol{\eta}_{ik})]^T \odot [\mathbf{1}\mathbf{t}_{ik}^T + \mathbf{t}_{ik}\mathbf{1}^T],\end{aligned}$$

where $\boldsymbol{V}(\boldsymbol{\gamma}_{ik})$ is given by (3). When $\tau_i^2 = 0$ the REMII model variance matrix (12) reduces to the REMI model (11). In the absence of overdispersion, that is with $\nu_i^2 = 0$ and $\tau_i^2 = 0$, the variance matrix once again reduces to the cumulative multinomial variance structure given by (3).

Here we simplify matters by assuming uncorrelated random effects, taking $\lambda_i = 0$; a reasonable assumption when working with the centered version of time as the covariate. Hence we only need to estimate the variances ν_i^2 and τ_i^2.

4. Estimation

Here we outline the parameter estimation procedures for the models discussed in Section 3. For the *B. bassiana* bioassay, because of the equal sample sizes, we have a simple constant overdispersion factor for the DM model. Standard QL techniques (McCullagh & Nelder, 1989) give estimates of the vector of regression coefficients and moment methods can be used to estimate the overdispersion parameter.

The GEEs for the vector of regression coefficients (Liang & Zeger, 1986) are given by

$$\begin{aligned}\mathbf{U}(\boldsymbol{\beta}_i) &= \sum_{k=1}^{K}\left[\frac{\partial\boldsymbol{\gamma}_{ik}}{\partial\boldsymbol{\beta}_i^T}\right][\mathrm{Var}(\boldsymbol{R}_{ik})]^{-1}(\boldsymbol{r}_{ik} - \boldsymbol{\gamma}_{ik}) \\ &= \sum_{k=1}^{K}\boldsymbol{X}_{ik}^T\boldsymbol{\Delta}_{ik}^{-1}[\boldsymbol{V}_{ik}^{\mathrm{OD}}]^{-1}(\boldsymbol{r}_{ik} - \boldsymbol{\gamma}_{ik}) = 0, \qquad (13)\end{aligned}$$

where $\boldsymbol{X}_{ik}$ is a $D \times m$ matrix of covariates that includes functions of time, $\boldsymbol{\Delta}_{ik} = \mathrm{diag}\{\partial\eta_{ikj}/\partial\gamma_{ikj}\}$ is a $D \times D$ diagonal matrix, $\boldsymbol{V}_{ik}^{\mathrm{OD}}$ is given by (9), (11) or (12) and $\boldsymbol{r}_{ik}$ is the $D \times 1$ vector of observed cumulative proportions of insects dead by the jth day. Even when $\boldsymbol{V}_{ik}^{\mathrm{OD}}$ is misspecified, $\mathrm{E}[\mathbf{U}(\boldsymbol{\beta}_i)] = 0$ and hence the QL estimate (13) is consistent.

Fixing ρ_i, σ_i^2 or $\{\nu_i^2, \tau_i^2\}$, the estimate $\widehat{\boldsymbol{\beta}}_i$ is found by an iterative weighted least squares procedure, where the iterative equations for $\boldsymbol{\beta}_i$ are given by

$$\widehat{\boldsymbol{\beta}}_i^{(s+1)} = \left(\sum_{k=1}^{K}\boldsymbol{X}_{ik}^T\boldsymbol{W}_{ik}^{\mathrm{OD}}\boldsymbol{X}_{ik}\right)^{-1}\sum_{k=1}^{K}\boldsymbol{X}_{ik}^T\boldsymbol{W}_{ik}^{\mathrm{OD}}\boldsymbol{z}_{ik}, \qquad (14)$$

with

$$\boldsymbol{W}_{ik}^{\mathrm{OD}} = \boldsymbol{\Delta}_{ik}^{-1}\left[\boldsymbol{V}_{ik}^{\mathrm{OD}}\right]^{-1}\boldsymbol{\Delta}_{ik}^{-1},$$

where $\boldsymbol{z}_{ik} = \boldsymbol{X}_{ik}\widehat{\boldsymbol{\beta}}_i^{(s)} + \boldsymbol{\Delta}_{ik}(\boldsymbol{r}_{ik} - \boldsymbol{\gamma}_{ik})$ with both $\boldsymbol{\Delta}_{ik}$ and $\boldsymbol{\gamma}_{ik}$ evaluated at the current estimate $\widehat{\boldsymbol{\beta}}_i^{(s)}$. The estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_i$ is given by

$$\mathbf{M}_0 = \left(\sum_{k=1}^{K}\boldsymbol{X}_{ik}^T\boldsymbol{\Delta}_{ik}^{-1}[\boldsymbol{V}_{ik}^{\mathrm{OD}}]^{-1}\boldsymbol{\Delta}_{ik}^{-1}\boldsymbol{X}_{ik}\right)^{-1}. \qquad (15)$$

In Liang & Zeger (1986) it is shown that under mild regularity conditions, $\widehat{\boldsymbol{\beta}}_i$, the solution of (13), is consistent and asymptotically multivariate Gaussian with covariance matrix given by the sandwich estimator $\mathbf{S} = \mathbf{M}_0\mathbf{M}_1^{-1}\mathbf{M}_0$, where $\mathbf{M}_0$ is given by (15) and

$$\mathbf{M}_1 = \left(\sum_{k=1}^{K} \boldsymbol{X}_{ik}^T \boldsymbol{\Delta}_{ik}^{-1} [\boldsymbol{V}_{ik}^{\text{OD}}]^{-1} (\boldsymbol{r}_{ik} - \widehat{\boldsymbol{\gamma}}_{ik})(\boldsymbol{r}_{ik} - \widehat{\boldsymbol{\gamma}}_{ik})^T [\boldsymbol{V}_{ik}^{\text{OD}}]^{-1} \boldsymbol{\Delta}_{ik}^{-1} \boldsymbol{X}_{ik} \right)^{-1}.$$

Note that this is robust to the choice of $\boldsymbol{V}_{ik}^{\text{OD}}$.

For the DM model we use a moment method (Moore, 1987; Williams, 1982) to estimate the overdispersion parameter in (9). Equating the Pearson's chi-square statistic X^2 to its expected value (solving the equation $X^2 = \text{E}(X^2)$) gives an estimating equation for ρ_i

$$\sum_{k=1}^{K} \frac{(\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik})^T \boldsymbol{V}^{-1}(\hat{\boldsymbol{\gamma}}_{ik})(\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik})}{1 + (n_{ik} - 1)\widehat{\rho}_i} = KD - m$$

and for $n_{ik} = n$ the estimate of the overdispersion parameter is given by

$$\widehat{\rho}_i = \frac{1}{n-1} \left(\sum_{k=1}^{K} \frac{(\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik})^T \boldsymbol{V}^{-1}(\hat{\boldsymbol{\gamma}}_{ik})(\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik})}{KD - m} - 1 \right). \tag{16}$$

In general we could iterate between solving (16) and estimating $\boldsymbol{\beta}_i$ from (14), with updated $\boldsymbol{z}_{ik}$ and $\boldsymbol{W}_{ik}^{\text{OD}}$, until convergence. However here, since $\boldsymbol{V}^{\text{OD}}$ for the DM model is only a constant multiple of the base multinomial variance matrix $\boldsymbol{V}$, the $\boldsymbol{\beta}_i$ estimate is identical to that from the multinomial model. So we simply need to use these estimates and solve (16) to find the $\widehat{\rho}_i$; there is no need for iteration.

For the REMI model (11) using the moment method to estimate σ_i^2 leads to the following iterative estimating equation where at the $(s+1)$-th step

$$\sigma_i^2(s+1) = \frac{1}{K-1} \sum_{k=1}^{K} \frac{\sigma_i^2(s)[\mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik})^T(\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik})]^2}{\mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik})^T \boldsymbol{V}_{ik}^I(s) \mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik})},$$

where $\boldsymbol{V}_{ik}^I(s)$ is given by (11) evaluated at $\sigma_i^2(s)$ the current estimate of σ_i^2. However, as the linear predictor is identical for each replicate we have $\eta_{ik} = \eta_i$ and also V_{ik}^I only depends on the isolate i, the above estimating equation simplifies and leads to a closed form solution

$$\widehat{\sigma}_i^2 = \frac{\dfrac{1}{K-1} \displaystyle\sum_{k=1}^{K} \left[\mathbf{h}'(\hat{\boldsymbol{\eta}}_i)^T (\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik}) \right]^2 - \mathbf{h}'(\hat{\boldsymbol{\eta}}_i)^T \boldsymbol{V}_i \mathbf{h}'(\hat{\boldsymbol{\eta}}_i)}{\left(1 - \dfrac{1}{n}\right) \left[\mathbf{h}'(\hat{\boldsymbol{\eta}}_i)^T \mathbf{h}'(\hat{\boldsymbol{\eta}}_i) \right]^2}.$$

Similarly, to estimate ν_i^2 and τ_i^2 in the REMII model (12) the iterative equations are

$$\nu_i^2(s+1) = \frac{1}{K-1} \sum_{k=1}^{K} \frac{\nu_i^2(s)[\mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik})^T(\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik})]^2}{\mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik})^T \boldsymbol{V}_{ik}^{II}(s) \mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik})} \tag{17}$$

and

$$\tau_i^2(s+1) = \frac{1}{K-1} \sum_{k=1}^{K} \frac{\tau_i^2(s)[(\mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik}) * \mathbf{t}_{ik})^T(\boldsymbol{r}_{ik} - \hat{\boldsymbol{\gamma}}_{ik})]^2}{[\mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik}) * \mathbf{t}_{ik}]^T \boldsymbol{V}_{ik}^{II}(s) [\mathbf{h}'(\hat{\boldsymbol{\eta}}_{ik}) * \mathbf{t}_{ik}]}, \tag{18}$$

where $\boldsymbol{V}_{ik}^{\mathrm{II}}(s)$ is given by (12) evaluated at the current estimates $\nu_i^2(s)$ and $\tau_i^2(s)$. Again, taking advantage of the model simplifications gives a simple iterative scheme for updating the variance parameters at each step of the overall estimation process. Moreover, because of the simple regular structure here, after some algebra solving the joint estimating equations for ν_i^2 and τ_i^2 in (17) and (18), we are able to obtain closed form estimates for these random effect variance, see Appendix 1 for details.

5. Results

Fitting the four models described in Section 3 to the 30 isolates shown in Figure 1 we find that for each isolate the estimated regression parameters (β_{1i} and β_{2i}) are identical. That the multinomial and Dirichlet-multinomial models give identical estimates is to be expected because of the proportional variance matrices, although it is somewhat surprising that it should also be true for the REMI and REMII models and this aspect is explored in Section 6. With the identical parameter estimates the fitted responses (as shown in Figure 1) are also identical, as are the estimated $LT50$s. Where the models do differ is in the estimated overdispersion parameters and resulting asymptotic covariance matrices and, hence, in the estimated asymptotic standard errors for the regression parameters, the fitted curves, and the $LT50$s. However, it turns out that the robust sandwich estimator covariance matrices are also identical over the four models; a property that is also explored in Section 6.

Table 1. Summary Statistics for estimated overdispersion parameters of DM, REMI and REMII models, in which the overdispersion is allowed to vary over isolates, and estimated $LT50$.

Model	Overdispersion	Estimate			
	Parameter	Mean	Median	Min.	Max.
DM	ρ_i	0.1156	0.1227	0.0156	0.2996
REMI	σ_i^2	1.0338	0.7294	0.0874	3.0359
REMII	ν_i^2	0.8728	0.6096	–0.4268	3.0359
	τ_i^2	0.0615	0.0328	–0.0513	0.3867
$LT50$		7.3007	6.7297	4.1743	11.7718

In Table 1 we present summary statistics for the various overdispersion parameters for the Dirichlet-multinomial, REMI, and REMII models. These parameters reflect a combination of the replicate variability and the lack of fit of the fitted responses to the mean response for each isolate. What is clear from this table is the wide range of values reflecting what is seen in Figure 1; some of the isolates exhibit very little variation over the replicates, while others are very disparate. The other notable aspect in this table is that from the REMII model some of the isolates lead to negative values of ν^2 and/or τ^2. This of course contradicts the conditional model formulation where these are interpreted as variances. However, with the GEE fitting approach adopted here these are really only additional parameters in an extended variance function and they are not constrained to be positive; the only constraint is that the resulting overdispersed variance matrix, $\boldsymbol{V}^{OD}$, should be positive definite.

As noted in the original discussion of the data, a simple aim of the study is to select effective isolates that are reliable with low variation across replicates. A simple summary of the efficacy is the estimated $LT50$ (identical across all models) and the estimated overdispersion parameters provide a simple and useful summary of the variability. For illustration we use the σ_i^2 from the REMI model. Figure 2 shows a plot of these quantities for the 30 isolates under consideration. The points are plotted with the isolate codes, which allows reference back to the raw data displayed in Figure 1. Focussing on those isolates with an estimated $LT50$ <6 (based on the range of values in Table 1) and $\hat{\sigma}_i^2 < 1$ we see that these isolates (602, 738, 743, 787, 841, 845, 848, 1028) do indeed correspond to

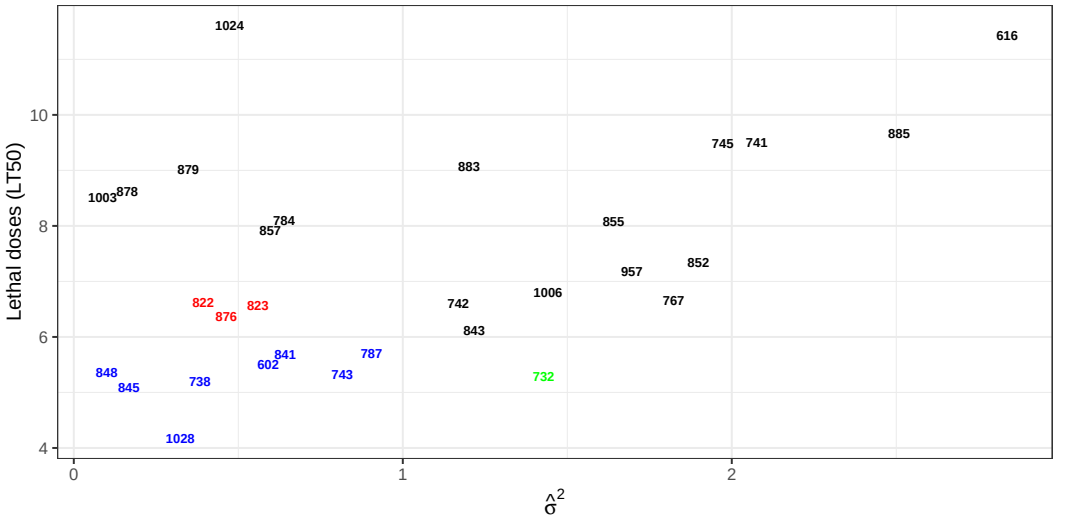

Figure 2. REMI fitted model: estimated $LT50$ values for isolates versus estimated random effect variances $\hat{\sigma}^2$. The colours indicate different groups of isolates, with blue corresponding to the effective, consistent ones.

those which are effective and relatively consistent. There is a single isolate (732) which is effective with a small $LT50$, but has relatively large variability. There is also a further group of three isolates (822, 823, 876) with low variability but $LT50$s slightly greater than 6. Of course, the setting of thresholds, as here, is somewhat arbitrary but it does give a coherent method for selecting and screening isolates and is easily applied to larger datasets, such as the full set of 142 isolates.

All computations reported in this paper were carried out in R (R Development Core Team, 2010). Data and code will be made available in due course.

6. Theoretical discussion of results

In the previous Section we noted that the linear predictor parameter estimates from all of the models are ***identical***. It is not surprising that they are similar as generalized least-squares estimates are not very sensitive to the precise form of the variance (weight) matrix, however that they are identical is more notable. The second surprising outcome is that the robust standard errors are also ***identical***. The same behaviour is found with the data set from O'Hara Hines & Lawless (1993) when excluding all the covariates from the model and fitting just time. Examination of the model based (asymptotic) standard errors suggested that there were also simple relationships between these across the different models. There is clearly something special going on here and this was found to depend on the special regular form of the data and certain aspects of the different models being considered. Key to understanding this are the mathematical results stated in the next section (with associated proofs in Appendix 2). In its simplest form this is related to the results in McElroy (1976), Rao's special form of covariance matrix (Rao, 1967) and also the work of Lange & Laird (1989), who consider balanced growth curve models for normally distributed responses. Here the details are somewhat different as we are working in the iterative framework of GEE for non-normal responses.

6.1 A Matrix Result

For each isolate, the model fitting used in this paper is an iterative application of generalized least-squares where for a variance matrix $\boldsymbol{V}$ we have

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\mathbf{z}.$$

If we now consider fitting variants of this with a modified variance function

$$\widetilde{\boldsymbol{V}} = \boldsymbol{V} + \boldsymbol{X}\mathbf{C}\boldsymbol{X}^T$$

for some specified matrix $\mathbf{C}$, then the resulting parameter estimate is given by

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\widetilde{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widetilde{\boldsymbol{V}}^{-1}\mathbf{z}.$$

The, somewhat, surprising results are:

- the parameter estimates from the two fits are the identical, i.e. $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$;
- the model-based asymptotic covariance matrix

$$\mathbf{M}_0 = \left(\boldsymbol{X}^T\mathbf{W}\boldsymbol{X}\right)^{-1}$$

  and that for the modified variance matrix $\widetilde{\boldsymbol{V}}$

$$\widetilde{\mathbf{M}}_0 = \left(\boldsymbol{X}^T\widetilde{\mathbf{W}}\boldsymbol{X}\right)^{-1}$$

  are simply related with

$$\widetilde{\mathbf{M}}_0 = \mathbf{M}_0 + \mathbf{C};$$

- the robust sandwich variance matrices $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ are identical.

We now consider how these results apply to the various models considered here and explain the observed relationships between the various fits.

6.2 Application to GEE for Cumulative Multinomial

The first simplification here is that every isolate has the same set of covariates and so we can write $\boldsymbol{X}_{ik} = \boldsymbol{X}$ and further within each isolate the replicates have identical linear predictors, that is $\gamma_{ik} = \gamma_i$ for $k = 1, \ldots, K$. Hence, the score function for $\boldsymbol{\beta}_i$ can be written as

$$\mathbf{U}(\boldsymbol{\beta}_i) = \sum_{k=1}^{K} \boldsymbol{X}^T\boldsymbol{\Delta}_i^{-1}[\boldsymbol{V}_i^{\text{OD}}]^{-1}(\boldsymbol{r}_{ik} - \boldsymbol{\gamma}_i) = K\boldsymbol{X}^T\boldsymbol{\Delta}_i^{-1}[\boldsymbol{V}_i^{\text{OD}}]^{-1}(\bar{\boldsymbol{r}}_i - \boldsymbol{\gamma}_i)$$

where $\bar{\boldsymbol{r}}_i$ are the mean cumulative proportions for the ith isolate. The iterative estimating equation (14) for $\boldsymbol{\beta}_i$ can then be written as

$$\widehat{\boldsymbol{\beta}}_i^{(s+1)} = \left(K\boldsymbol{X}^T\boldsymbol{W}_i^{\text{OD}}\boldsymbol{X}\right)^{-1} K\boldsymbol{X}^T\boldsymbol{W}_i^{\text{OD}}\bar{\boldsymbol{z}}_i = \left(\boldsymbol{X}^T\boldsymbol{W}_i^{\text{OD}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}_i^{\text{OD}}\bar{\boldsymbol{z}}_i,$$

where $\bar{\boldsymbol{z}}_i = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_i^{(s)} + \boldsymbol{\Delta}_i(\bar{\boldsymbol{r}}_i - \boldsymbol{\gamma}_i)$ and $\boldsymbol{W}_{ik}^{\text{OD}} = \boldsymbol{W}_i^{\text{OD}} = \boldsymbol{\Delta}_i^{-1}\left[\boldsymbol{V}_i^{\text{OD}}\right]^{-1}\boldsymbol{\Delta}_i^{-1}$. Note that $\bar{\boldsymbol{z}}_i$ only depends upon the current parameter estimates, the data (response and explanatory variables) and the link function.

The estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_i$ is given by

$$\mathbf{M}_{0,i}^{OD} = \left(\sum_{k=1}^{K} \boldsymbol{X}^T \boldsymbol{\Delta}_i^{-1} [\boldsymbol{V}_i^{\text{OD}}]^{-1} \boldsymbol{\Delta}_i^{-1} \boldsymbol{X} \right)^{-1} = \frac{1}{K} \left(\boldsymbol{X}^T \boldsymbol{W}_i^{\text{OD}} \boldsymbol{X} \right)^{-1}$$

and the sandwich filling reduces to

$$\begin{aligned} \mathbf{M}_{1,i}^{OD} &= \left(\sum_{k=1}^{K} \boldsymbol{X}^T \boldsymbol{\Delta}_i^{-1} [\boldsymbol{V}_i^{\text{OD}}]^{-1} (\boldsymbol{r}_{ik} - \widehat{\boldsymbol{\gamma}}_i)(\boldsymbol{r}_{ik} - \widehat{\boldsymbol{\gamma}}_i)^T [\boldsymbol{V}_i^{\text{OD}}]^{-1} \boldsymbol{\Delta}_i^{-1} \boldsymbol{X} \right)^{-1} \\ &= \left(\boldsymbol{X}^T \boldsymbol{\Delta}_i^{-1} [\boldsymbol{V}_i^{\text{OD}}]^{-1} \left\{ \sum_{k=1}^{K} (\boldsymbol{r}_{ik} - \widehat{\boldsymbol{\gamma}}_i)(\boldsymbol{r}_{ik} - \widehat{\boldsymbol{\gamma}}_i)^T \right\} [\boldsymbol{V}_i^{\text{OD}}]^{-1} \boldsymbol{\Delta}_i^{-1} \boldsymbol{X} \right)^{-1}. \end{aligned}$$

6.3 DM

For the Dirichlet - Multinomial model $\boldsymbol{V}_i^{OD}$ is simply a scaled version of $\boldsymbol{V}(\boldsymbol{\gamma}_i)$ with multiplicative scale factor $1 + \rho_i(n-1)$ and so we have $\mathbf{W}_i^{OD} = \mathbf{W}_i/[1 + \rho_i(n-1)]$. Hence the ρ_i cancels in estimating equation for $\boldsymbol{\beta}_i$ giving identical parameter estimates as for the basic multinomial model. The estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_i$ is scaled by $1 + \rho_i(n-1)$ giving standard errors that are inflated by a factor of $\sqrt{1 + \rho_i(n-1)}$. However, in calculating the robust standard errors the scaling factor cancels in the sandwich estimator expression for Cov_i^* so the robust standard errors are identical to those from the multinomial model, which is consistent with the notion of these being in some sense robust to the choice of variance function.

6.4 REM models

For isolate i and the REM models we have

$$\boldsymbol{V}_i^{OD} = \boldsymbol{V}_i + \text{additional terms},$$

where $\boldsymbol{V}_i = \boldsymbol{V}(\boldsymbol{\gamma}_i)$ is the variance matrix for the basic cumulative multinomial model. Subsequently, for REMI and REMII we will show that these are in the special form required to apply the matrix results from Section 6.1. In particular we can express these as

$$\boldsymbol{\Delta}_i \boldsymbol{V}_i^{\text{OD}} \boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i \boldsymbol{V}_i \boldsymbol{\Delta}_i + \boldsymbol{X} \mathbf{C}_i \boldsymbol{X}^T,$$

where $\mathbf{C}_i$ depends on the particular approximate random effects model and the isolate. When this relationship holds, for a common $\mathbf{z}_i$ the variance matrices $\boldsymbol{V}_i$ and $\boldsymbol{V}_i^{\text{OD}}$ will give the same parameter estimate updates and so will define the same iterative path and converge to identical parameter estimates $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\boldsymbol{\beta}}_i^{\text{OD}}$. Further, because of the regular replicate structure, the asymptotic covariance matrices

$$\mathbf{M}_{0,i} = \frac{1}{K} \left(\boldsymbol{X}^T \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \quad \text{and} \quad \mathbf{M}_{0,i}^{OD} = \frac{1}{K} \left(\boldsymbol{X}^T \boldsymbol{W}_i^{\text{OD}} \boldsymbol{X} \right)^{-1}$$

will be related by

$$\mathbf{M}_{0,i}^{\text{OD}} = \mathbf{M}_{0,i} + \frac{\mathbf{C}_i}{K}.$$

Finally, the robust (sandwich) covariance matrix estimates will be identical

$$\mathbf{S}_i = \mathbf{M}_{0,i} \mathbf{M}_{1,i}^{-1} \mathbf{M}_{0,i} = \mathbf{M}_{0,i}^{\text{OD}} \left(\mathbf{M}_{1,i}^{\text{OD}} \right)^{-1} \mathbf{M}_{0,i}^{\text{OD}} = \mathbf{S}_i^{\text{OD}}.$$

6.4.1 REMI

Here we have

$$\boldsymbol{V}_i^I = \boldsymbol{V}_i + \left(1 - \frac{1}{n}\right)\sigma_i^2 \mathbf{h}_i' \mathbf{h}_i'^T,$$

where $\mathbf{h}_i' = \partial \mathbf{h}_i / \partial \boldsymbol{\eta}_i$. In the iterative estimation procedure we have $\Delta_i = \text{diag}(\partial \boldsymbol{\eta}_i / \partial / \boldsymbol{\gamma}_i)$ so $\Delta_i^{-1} = \text{diag}(\partial \boldsymbol{\gamma}_i / \partial / \boldsymbol{\eta}_i) = \text{diag}(\mathbf{h}_i')$ giving $\mathbf{h}_i' = \Delta_i^{-1}\mathbf{1}$ and $\Delta_i \mathbf{h}_i' = \mathbf{1}$. Hence,

$$\begin{aligned}\Delta_i \boldsymbol{V}_i^I \Delta_i &= \Delta_i \boldsymbol{V}_i \Delta_i + \left(1 - \frac{1}{n}\right)\sigma_i^2 \Delta_i \mathbf{h}_i' \mathbf{h}_i'^T \Delta_i \\ &= \Delta_i \boldsymbol{V}_i \Delta_i + \left(1 - \frac{1}{n}\right)\sigma_i^2 \mathbf{1}\mathbf{1}^T\end{aligned}$$

and noting that here $\boldsymbol{X}$ has the simple form $\boldsymbol{X} = [\mathbf{1}, \mathbf{t}]$ for all isolates, we can write

$$\Delta_i \boldsymbol{V}_i^I \Delta_i = \Delta_i \boldsymbol{V}_i \Delta_i + \boldsymbol{X}\mathbf{C}_i^I \boldsymbol{X}^T,$$

where

$$\mathbf{C}_i^I = \begin{bmatrix} \left(1 - \frac{1}{n}\right)\sigma_i^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence, we have the modified iterative variance function $\Delta_i \boldsymbol{V}_i^{OD} \Delta_i$ of precisely the form to apply the results of Section 6.1, with the $\mathbf{C}$ matrix having a very simple form.

So all of the previously discussed results will apply: the parameter estimates will be identical to those from the basic multinomial model; the robust sandwich covariance matrices, and hence robust standard errors, will be identical; finally, taking account of the $K = 5$ replicates, the model based asymptotic covariance matrices will be related through

$$\mathbf{M}_0^I = \mathbf{M}_0 + \frac{\mathbf{C}_i^I}{5} = \mathbf{M}_0 + \frac{1}{5}\begin{bmatrix} \left(1 - \frac{1}{n}\right)\sigma_i^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

So the only impact is on the variance estimate (standard error) for the intercept term $\beta_{0,i}$ and we have

$$\text{se}^I(\widehat{\beta}_{0,i}) = \sqrt{\text{se}(\widehat{\beta}_{0,i})^2 + \frac{1}{5}\left(1 - \frac{1}{n}\right)\widehat{\sigma}_i^2}.$$

For illustration, we consider isolate 848 where $\text{se}(\widehat{\beta}_{1,i}) = 0.1452$, $\text{se}^I(\widehat{\beta}_{1,i}) = 0.2012$, and $\widehat{\sigma}^2 = 0.1002$ and we have

$$\sqrt{\text{se}(\widehat{\beta}_{1,i})^2 + \frac{1}{5}\left(1 - \frac{1}{n}\right)\widehat{\sigma}_i^2} = \sqrt{0.1452^2 + \frac{1}{5}\left(1 - \frac{1}{30}\right)0.1002} = 0.2012,$$

giving the expected agreement. In addition, the standard error for the time coefficient $\beta_{1,i}$ is unchanged at 0.0378.

6.4.2 REMII

Using the mean-centered version of time and assuming uncorrelated random effects for the slopes and intercept (i.e. $\lambda_i = 0$) the REMII variance function can be expressed as

$$\boldsymbol{V}_i^{II} = \boldsymbol{V}_i + \left(1 - \frac{1}{n}\right)\left\{\nu_i^2 \mathbf{h}_i' \mathbf{h}_i'^T + \tau_i^2 [\mathbf{h}' \odot \mathbf{t}][\mathbf{h}' \odot \mathbf{t}]^T\right\}.$$

In the same way as above we can now write

$$\begin{aligned}\Delta_i \boldsymbol{V}_i^{II} \Delta_i &= \Delta_i \boldsymbol{V}_i \Delta_i + \left(1 - \frac{1}{n}\right)\left\{\nu_i^2 \Delta_i \mathbf{h}_i' \mathbf{h}_i'^T \Delta_i + \tau_i^2 [\Delta_i \mathbf{h}' \odot \mathbf{t}][\Delta_i \mathbf{h}' \odot \mathbf{t}]^T\right\} \\ &= \Delta_i \boldsymbol{V}_i \Delta_i + \left(1 - \frac{1}{n}\right)\left\{\nu_i^2 \mathbf{1}\mathbf{1}^T + \tau_i^2 [\mathbf{1} \odot \mathbf{t}][\mathbf{1} \odot \mathbf{t}]^T\right\} \\ &= \Delta_i \boldsymbol{V}_i \Delta_i + \left(1 - \frac{1}{n}\right)\left\{\nu_i^2 \mathbf{1}\mathbf{1}^T + \tau_i^2 \mathbf{t}\mathbf{t}^T\right\}.\end{aligned}$$

Again noting that $\boldsymbol{X} = [\mathbf{1}, \mathbf{t}]$ we can write the above as

$$\Delta_i \boldsymbol{V}_i^{II} \Delta_i = \Delta_i \boldsymbol{V}_i \Delta_i + \boldsymbol{X} \mathbf{C}_i^{II} \boldsymbol{X}^T,$$

where

$$\mathbf{C}_i^{II} = \left(1 - \frac{1}{n}\right)\begin{bmatrix} \nu_i^2 & 0 \\ 0 & \tau_i^2 \end{bmatrix}.$$

Then from the general results, as with REMI, the parameter estimates and robust standard errors are unchanged and the asymptotic covariance matrices are related by

$$\mathbf{M}_0^{II} = \mathbf{M}_0 + \frac{\mathbf{C}_i^{II}}{5} = \mathbf{M}_0 + \frac{1}{5}\left(1 - \frac{1}{n}\right)\begin{bmatrix} \nu_i^2 & 0 \\ 0 & \tau_i^2 \end{bmatrix},$$

with the respective variance parameter estimates modifying the standard errors in quadrature. For illustration with isolate 848 $\hat{\nu}_i^2 = 0.0968$ and $\hat{\tau}_i^2 = 0.0121$ giving

$$\mathrm{se}^{II}(\widehat{\beta}_{1,i}) = \sqrt{\mathrm{se}(\widehat{\beta}_{1,i})^2 + \frac{1}{5}\left(1 - \frac{1}{n}\right)\widehat{\nu}_i^2} = \sqrt{0.1452^2 + \frac{1}{5}\left(1 - \frac{1}{30}\right)0.0968} = 0.1995$$

and

$$\mathrm{se}^{II}(\widehat{\beta}_{2,i}) = \sqrt{\mathrm{se}(\widehat{\beta}_{2,i})^2 + \frac{1}{5}\left(1 - \frac{1}{n}\right)\widehat{\tau}_i^2} = \sqrt{0.0378^2 + \frac{1}{5}\left(1 - \frac{1}{30}\right)0.0121} = 0.0614,$$

both of which are in agreement with the results from the iterative fitting process.

7. Final Remarks

This paper focusses on a specific set of overdispersed multinomial models for what is essentially longitudinal grouped mortality data. There are many other related approaches that can be taken for data of this form. In a forthcoming paper, Fallah *et al.* (2023), we consider and compare the use of ordinal data models (as in Martinez & Hinde, 2014) and discrete survival models again incorporating random effects, but for both replicates and isolates. In addition, we consider the use of arbitrary random effects for isolates using non-parametric maximum likelihood estimation (NPMLE), see Einbeck *et al.* (2018), which allows for implicit clustering of the isolates and a more systematic approach to the selection of the best, rather than the more *ad hoc* approach adopted in Section 5.

Another open topic is that of model fit and model diagnostics. O'Hara Hines *et al.* (1992) consider these aspects for fitting the basic multinomial model to similar grouped cumulative mortality data. Because of the replication it is possible to decompose an overall goodness of fit measure, such as

deviance or generalized Pearson statistic, into a lack of fit term based on the mean proportions over the replicates and a term capturing the replicate variability. The first terms allows a comparison of the fit from alternative linear predictors, as in Martinez & Hinde (2014), and it would, in principle, be possible to estimate overdispersion parameters and random effect variances from the replicate variability. Here, we took a simpler approach basing overdispersion parameter estimation on the overall residual lack of fit, as is common in applications of overdispesion models. Diagnostic assessment of model fit and the adequacy of specific overdispersion models could also, in principle, be explored using half-normal plots as in Salvador *et al.* (2022) for standard multinomial and Dirichlet-multinomial models. To apply such methods here would require extending the approach to the cumulative type of response considered here through, for example, providing these new model types for the `hnp` package (Moral *et al.,* 2017) in R.

Acknowledgements

The authors are grateful to the late Prof. Sérgio Batista, from São Paulo University (ESALQ/USP) for supplying the data. The first author Silvia M. de Freitas acknowledges the scholarship financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. This research was partially supported by FAPESP Project: 2022/11865-1, Brazil for John Hinde and by CNPq Project: 312645/2021-0, Brazil for Clarice G. B. Demétrio.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix 1: REMII estimates

For the REMII model taking advantage of the model simplifications gives a simple iterative scheme for updating the variance parameters at each step of the overall estimation process. Because of the simple regular structure here, after some algebra solving the joint estimating equations for ν_i^2 and τ_i^2 in (17) and (18), we are able to obtain closed form estimates for these random effect variances. While written in full these have a complicated appearance, by defining some key expressions these

can be written in a fairly simple overall form. We first define the key components

$$
\begin{aligned}
SS_h &= \frac{1}{K-1}\sum_{k=1}^{K}\left[\mathbf{h}'(\boldsymbol{\eta}_{ik})^T(\boldsymbol{r}_{ik}-\boldsymbol{\gamma}_{ik})\right]^2\\
SS_{ht} &= \frac{1}{K-1}\sum_{k=1}^{K}\left[\left(\mathbf{h}'(\boldsymbol{\eta}_{ik})^T * \mathbf{t}\right)(\boldsymbol{r}_{ik}-\boldsymbol{\gamma}_{ik})\right]^2\\
H &= \left[\mathbf{h}'(\boldsymbol{\eta}_{ik})^T\mathbf{h}'(\boldsymbol{\eta}_{ik})\right]^2\\
H_{tt} &= \left[\left(\mathbf{h}'(\boldsymbol{\eta}_{ik})^T * \mathbf{t}\right)^T\left(\mathbf{h}'(\boldsymbol{\eta}_{ik})^T * \mathbf{t}\right)\right]^2\\
H_t &= \left[\left(\mathbf{h}'(\boldsymbol{\eta}_{ik})^T * \mathbf{t}\right)^T)\mathbf{h}'(\boldsymbol{\eta}_{ik})\right]^2\\
V_h &= \mathbf{h}'(\boldsymbol{\eta}_{ik})^T V_{ik}\mathbf{h}'(\boldsymbol{\eta}_{ik})\\
V_{h,tt} &= \left(\mathbf{h}'(\boldsymbol{\eta}_{ik})^T * \mathbf{t}\right)^T V_{ik}\left(\mathbf{h}'(\boldsymbol{\eta}_{ik})^T * \mathbf{t}\right)
\end{aligned}
$$

$$
\nu_i^2 = \frac{H_{tt}(SS_h - V_h) - H_t SS_{ht}}{\left(1-\frac{1}{n}\right)\{HH_{tt} - H_t^2\}}
$$

and

$$
\tau_i^2 = \frac{H(SS_{ht} - V_{h,tt}) - H_t SS_h}{\left(1-\frac{1}{n}\right)\{HH_{tt} - H_t^2\}}.
$$

Appendix 2: Proofs of theoretical matrix results

We begin by considering a generalized weighted least-squares fit of a working response vector $\mathbf{z}$

$$
\widehat{\boldsymbol{\beta}} = (X^T\mathbf{W}X)^{-1}X^T\mathbf{W}\mathbf{z},
$$

where the weight matrix $\mathbf{W}$ is based on a variance matrix $V = \mathbf{W}^{-1}$. We consider specific modifications of the variance matrix of the form

$$
\widetilde{V} = V + X\mathbf{C}X^T \tag{19}
$$

and are interested in how using the corresponding weight matrix $\widetilde{\mathbf{W}} = \widetilde{V}^{-1}$ affects the parameter estimates, the asymptotic covariance matrix and the robust sandwich covariance matrix.

Pre- and post-multiplying (19) by $\widetilde{V}^{-1}$ and V^{-1} respectively gives an expression linking the inverse variances

$$
\widetilde{V}^{-1} = V^{-1} - \widetilde{V}^{-1}X\mathbf{C}X^TV^{-1},
$$

or in terms of the weight matrices weights

$$
\widetilde{\mathbf{W}} = \mathbf{W} - \widetilde{\mathbf{W}}X\mathbf{C}X^T\mathbf{W}.
$$

Parameter estimates

We define the parameter estimate based on the modified variance matrix as $\widetilde{\boldsymbol{\beta}}$ with

$$\widetilde{\boldsymbol{\beta}} = (X^T\widetilde{\mathbf{W}}X)^{-1}X^T\widetilde{\mathbf{W}}\mathbf{z}.$$

We now consider how this differs from the original estimate $\widehat{\boldsymbol{\beta}}$

$$\begin{aligned}
(X^T\widetilde{\mathbf{W}}X)(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) &= X^T\widetilde{\mathbf{W}}\mathbf{z} - (X^T\widetilde{\mathbf{W}}X)\widehat{\boldsymbol{\beta}} \\
&= X^T\widetilde{\mathbf{W}}(\mathbf{z} - X\widehat{\boldsymbol{\beta}}) = X^T\widetilde{\mathbf{W}}\left(\mathbf{z} - X(X^T\mathbf{W}X)^{-1}X^T\mathbf{W}\mathbf{z}\right) \\
&= \left[X^T\widetilde{\mathbf{W}} - X^T\widetilde{\mathbf{W}}X(X^T\mathbf{W}X)^{-1}X^T\mathbf{W}\right]\mathbf{z} \\
&= \left\{X^T\widetilde{\mathbf{W}}\mathbf{W}^{-1} - X^T\widetilde{\mathbf{W}}X\left[(X^T\widetilde{\mathbf{W}}X)^{-1} - \mathbf{C}\right]X^T\right\}\mathbf{W}\mathbf{z} \\
&= \left[X^T\widetilde{\mathbf{W}}\mathbf{W}^{-1} - X^T + X^T\widetilde{\mathbf{W}}X\mathbf{C}X^T\right]\mathbf{W}\mathbf{z} \\
&= \left[X^T\widetilde{\mathbf{W}}\left(\mathbf{W}^{-1} + X\mathbf{C}X^T\right) - X^T\right]\mathbf{W}\mathbf{z} \\
&= \left[X^T\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^{-1} - X^T\right]\mathbf{W}\mathbf{z} \\
&= (X^T - X^T)\mathbf{W}\mathbf{z} = \mathbf{0} \\
\Rightarrow \widetilde{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}}.
\end{aligned}$$

So the parameter estimates are unchanged.

Note also that if this forms part of an iterative estimation procedure, as in iteratively weighted least-squares, then, given the same starting values, the two variance functions will lead to the same iterative sequence and converge to the same final estimates.

Asymptotic covariance matrices

The asymptotic covariance matrix for the original variance matrix is given by

$$\mathbf{M}_0 = \left(X^T\mathbf{W}X\right)^{-1}$$

and for the modified variance matrix $\widetilde{V}$ we define

$$\widetilde{\mathbf{M}}_0 = \left(X^T\widetilde{\mathbf{W}}X\right)^{-1}.$$

Again looking at the difference in these matrices we have

$$\begin{aligned}
\widetilde{\mathbf{M}}_0 - \mathbf{M}_0 &= \left(X^T\widetilde{\mathbf{W}}X\right)^{-1}\left[X^T\mathbf{W}X - X^T\widetilde{\mathbf{W}}X\right]\left(X^T\mathbf{W}X\right)^{-1} \\
&= \left(X^T\widetilde{\mathbf{W}}X\right)^{-1}\left[X^T\widetilde{\mathbf{W}}\left(\widetilde{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\right)\mathbf{W}X\right]\left(X^T\mathbf{W}X\right)^{-1} \\
&= \left(X^T\widetilde{\mathbf{W}}X\right)^{-1}\left(X^T\widetilde{\mathbf{W}}X\right)\mathbf{C}\left(X^T\mathbf{W}X\right)\left(X^T\mathbf{W}X\right)^{-1} \\
&= \mathbf{C}.
\end{aligned}$$

So we have a very simple relationship between the covariance matrices

$$\widetilde{\mathbf{M}}_0 = \mathbf{M}_0 + \mathbf{C}.$$

Robust (sandwich) covariance matrices

The two sandwich covariance matrix estimates can be written as

$$\mathbf{S} = \mathbf{M}_0 X^T \mathbf{W}\mathbf{R}\mathbf{W} X \mathbf{M}_0$$
$$\widetilde{\mathbf{S}} = \widetilde{\mathbf{M}}_0 X^T \widetilde{\mathbf{W}}\mathbf{R}\widetilde{\mathbf{W}} X \widetilde{\mathbf{M}}_0,$$

where $\mathbf{R}$ is a matrix of cross-products of residuals, which is identical for both settings as the parameter estimates, and hence fitted values, are the same.

Using the result that $\widetilde{\mathbf{W}} = \mathbf{W} - \widetilde{\mathbf{W}} X \mathbf{C} X^T \mathbf{W}$, we can write

$$\begin{aligned}\widetilde{\mathbf{S}} &= \widetilde{\mathbf{M}}_0 X^T \mathbf{W}\mathbf{R}\widetilde{\mathbf{W}} X \widetilde{\mathbf{M}}_0 - \widetilde{\mathbf{M}}_0 \widetilde{\mathbf{M}}_0^{-1} \mathbf{C} X^T \mathbf{W}\mathbf{R}\widetilde{\mathbf{W}} X \widetilde{\mathbf{M}}_0 \\ &= \mathbf{M}_0 X^T \mathbf{W}\mathbf{R}\widetilde{\mathbf{W}} X \widetilde{\mathbf{M}}_0 + \mathbf{C} X^T \mathbf{W}\mathbf{R}\widetilde{\mathbf{W}} X \widetilde{\mathbf{M}}_0 - \mathbf{C} X^T \mathbf{W}\mathbf{R}\widetilde{\mathbf{W}} X \widetilde{\mathbf{M}}_0 \\ &= \mathbf{M}_0 X^T \mathbf{W}\mathbf{R}\widetilde{\mathbf{W}} X \widetilde{\mathbf{M}}_0\end{aligned}$$

and proceeding in a similar fashion for the remaining $\widetilde{\mathbf{W}}$,

$$\widetilde{\mathbf{S}} = \mathbf{M}_0 X^T \mathbf{W}\mathbf{R}\mathbf{W} X \mathbf{M}_0 = \mathbf{S}.$$

The robust covariance matrices are ***identical*** under each specification.

References

1. Almeida, J. E. M., Alves, S. B. & Pereira, R. M. Selection of *Beauveria spp.* isolates for control of the termite *Heterotermes tenuis* (Hagen, 1958). *Journal of Applied Entomology* **121,** 539–543 (1997).
2. Corsini, N. & Viroli, C. Dealing with overdispersion in multivariate count data. *Computational Statistics & Data Analysis* **170,** 107447 (2022).
3. Crowder, M. Beta-binomial anova for proportions. *Applied Statistics* **27,** 34–37 (1978).
4. Einbeck, J., Darnell, R. & Hinde, J. `npmlreg`*: Nonparametric Maximum Likelihood Estimation for Random Effect Models* R package version 0.46-5 (2018). https://CRAN.R-project.org/package=npmlreg.
5. Fallah, L., Freitas, S. M. d., Demétrio, C. B. & Hinde, J. P. *Longitudinal modelling of grouped mortality data* In preparation. 2023.
6. Finney, D. *Probit Analysis. 3th ed.* (Cambridge University Press, London, 1971).
7. Freitas, S. M. d. *Modelos para Proporções com Superdispersão proveniente de Ensaios Toxicológicos no Tempo.* (Ph.D. thesis, ESALQ, Universidade de São Paulo, 2001).
8. Glonek, G. F. V. & McCullagh, P. Multivariate logistic models. *Journal of the Royal Statistical Society, Series B,* **57,** 149–192 (1995).
9. Hinde, J. P & Demétrio, C. G. B. Overdispersion: models and estimation. *Computational Statistics and Data Analysis* **27,** 151–170 (1998).
10. Lange, N. & Laird, N. M. The Effect of Covariance Structure on Variance Estimation in Balanced Growth-Curve Models with Random Parameters. *Journal of the American Statistical Association* **84,** 241–247 (1989).
11. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22 (1986).

12. Martinez, M.-J. & Hinde, J. *Random effects ordinal time models for grouped toxicological data from a biological control assay.* in *Statistical Modelling:Papers in Biostatistics and Bioinformatics* (eds MacKenzie, G. & Peng, D.) (Springer, 2014), 45–58.

13. McCullagh, P. Regression models for ordinal data (with Discussion). *Journal of the Royal Statistical Society, Series B,* **42,** 109–151 (1980).

14. McCullagh, P. & Nelder, J. A. *Generalized Linear Models. 2nd ed* (Chapman-Hall, London, 1989).

15. McElroy, F. W. Optimality of Least Squares in Linear Models with Unknown Error Covariance Matrix. *Journal of the American Statistical Association* **71,** 374–377 (1976).

16. Moore, D. F. Modelling the extraneous variance in the presence of extra-binomial variation. *Applied Statistics* **36,** 8–14 (1987).

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

18. Mosimann, J. E. On the compound multinomial distribution, the multivariate β–distribution, and correlations among proportions. *Biometrika* **49,** 65–82 (1962).

19. O'Hara Hines, R. J. & Lawless, J. F. Modelling overdispersion in toxicological mortality data grouped over time. *Biometrics* **49,** 107–121 (1993).

20. O'Hara Hines, R. J., Lawless, J. F. & Carter, E. M. Diagnostics for a Cumulative Multinomial Generalized Linear Model, with Applications to Grouped Toxicological Mortality Data. *Journal of the American Statistical Association* **87,** 1059–1069 (1992).

21. Petkau, A. J. & Sitter, R. R. Models for quantal response experiments over time. *Biometrics* **45,** 1299–1308 (1989).

22. R Development Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2010. http://www.r-project.org/.

23. Rao, C. R. *Least squares theory using an estimated dispersion matrix and its application to measurement of signals* in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* **5** (1967), 355–373.

24. Rao, C. R. *Linear Statistical Inference and its Applications* (John Wiley & Sons, Inc., 1973).

25. Salvador, M. L., Junior, E. E. R., Taconeli, C. A. & Lara, I. A. R. Analysis of multinomial data with overdispersion diagnostics and application. *Brazilian Journal of Biometrics* **40** (2022).

26. Tamai, M. A. *Avaliacão de fungos entomopatogenicos para o controle de Tetranychus urticae Koch* (Master thesis, ESALQ, Universidade de São Paulo, 1997).

27. Williams, D. A. Extra-binomial variation in logistic-linear models. *Applied Statistics* **31,** 144–148 (1982).