Models for overdispersed data in entomology

Models for overdispersed data in entomology

Clarice G.B. Demétrio, John Hinde and Rafael A. Moral

Abstract Entomological data are often overdispersed, characterised by a larger variance than assumed by simple standard models. It is important to model overdispersion properly in order to avoid incorrect and misleading inferences. Outcomes of interest are often in the form of counts or proportions and we present extended models that incorporate overdispersion, methods to assess its impact and model goodness-of-fit, and techniques to test treatment differences in the presence of overdispersion.

1 Introduction

Outcomes of interest for entomological data are often in the form of counts or proportions which can be analysed by 0. standard Poisson and binomial models, respectively. These are both specific examples of generalized linear models (McCullagh and Nelder, 1989) and hence our focus on this class of models. However, in general, the data are overdispersed, characterised by a larger variance than assumed by these simple standard models. It is important to model overdispersion properly in order to avoid incorrect and misleading inferences (Hinde and Demétrio, 1998).

There are many different possible causes of overdispersion and in any modeling situation a number of these could be involved. Some common possibilities are:

Clarice G.B. Demétrio
Escola Superior de Agricultura "Luiz de Queiroz ", Departamento de Cincias Exatas, Piracicaba, São Paulo, Brazil, e-mail: clarice.demetrio@usp.br

John Hinde
National University of Ireland, Galway, Ireland, e-mail: john.hinde@nuigalway.ie

Rafael Moral
Escola Superior de Agricultura "Luiz de Queiroz ", Departamento de Cincias Exatas, Piracicaba, São Paulo, Brazil, e-mail: rafael.moral@usp.br

1. Variability of experimental material — this can be thought of as individual variability of the experimental units and may give an additional component of variability that is not accounted for by the basic response model. For example, in dose-response experiments, the insects used will typically have differing susceptibilities to the substance that may affect the responses.
2. Correlation between individual responses – for example, in biological assays involving batches of insects we may expect to see some correlation between insects from the same batch since they may be genetically similar.
   Any context here for longitudinal observation?
3. Cluster and multistage sampling – here instead of a simple random sample we structure the population of insects into a hierarchy and sample sequentially from each level. For example, we may consider insects within metapopulations within ecosystems. In our sampling we may take a random sample of ecosystems, then from these selected ecosystems we may pick a random sample of metapopulations, and, finally, take our observational units from a random sample of insects in these selected metapopulations. This structured hierarchical sampling can lead to complex dependencies between the individual level responses and certainly we are likely to see correlation between the responses within a given metapopulation.
4. Aggregation – here the individual level responses are grouped into a response at a higher, aggregate, level. The aggregation process may be known, but more generally it is not completely specified and leads to a compound distribution for the observed responses. For example, in biological control studies we may observe total numbers of insects emerged from parasitised larvae by a number of females, but given that generally every female can lay a different number of eggs per host the observed totals of insects will be a combination of the number of females that parasitised the larvae and the distribution of the number of eggs per female; any modelling may more sensibly apply to the numbers of insects from the same female, but this is not observed, only the total number of insects emerged from a number of parasitised larvae.
5. Omitted unobserved variables – in some sense the other categories are all special cases of this, but generally in a rather complex way. Our models will often be formulated with the notion of some omitted variable to account for possible underlying, but unobserved, structure. This is particularly relevant in the regression modelling context.

In some circumstances the cause of the overdispersion may be apparent from the nature of the data collection process. Although, it should be noted that different explanations of the overdispersion process can lead to the same model. In general, it is difficult to infer the precise cause, or underlying process, leading to the overdispersion. However, the causes mentioned above provide a useful framework for thinking about overdispersion in practical applications, even if the distinctions are not always sharp. This will become apparent in the subsequent development of overdispersion models and some applications.

Overdispersion can arise in various ways, typically through some failure of the basic model (e.g. Poisson or binomial model) assumptions. In this chapter we will consider mechanisms that can lead to overdispersion for categorical and count data.

These mechanisms will suggest extensions to the basic model that can describe certain forms of overdispersed data. Using an inadequate model that does not account for overdispersion may lead us to make misleading overly precise inferences and predictions, as certainly standard errors will be incorrect and may be seriously underestimated. We present here extended models that incorporate overdispersion, methods to assess its impact and model goodness-of-fit, and techniques to test treatment differences in the presence of overdispersion.

2 Models for proportion data

2.1 Binomial - logit/probit/CLL models

For a group of m independent Bernoulli trials[1] with constant success probability π, the probability distribution for the total number of successes, Y, is binomial and

$$\Pr(Y = y) = \binom{m}{y} \pi^y (1-\pi)^{m-y} \quad y = 0, 1, \ldots, m \tag{1}$$

denoted by $\text{Bin}(m, \pi)$. This is simply the distribution of the sum of m independent Bernoulli random variables. The binomial coefficient $\binom{m}{y}$ arises from the number of possible sequences of m Bernoulli trials that give y successes.

If we suppose that the random variables Y_i, $i = 1, \ldots, n$, represent counts of successes out of samples of size m_i with $Y_i \sim \text{Bin}(m_i, \pi_i)$, then Y_i has mean

$$\text{E}(Y_i) = \mu_i = m_i \pi_i,$$

and variance

$$\text{Var}(Y_i) = m_i \pi_i (1-\pi_i) = \mu_i \left(1 - \frac{\mu_i}{m_i}\right). \tag{2}$$

It is important to note that the variance of Y_i is a simple fixed function of the mean which constrains how the model can account for the observed variability in proportion data. In general, for real data sets the observed variance is larger than that implied by the binomial model.

A generalized linear model allows us to model the expected proportions π_i in terms of explanatory variables $\mathbf{x}_i$ through a transformed llinear function

$$g(\pi_i) = \boldsymbol{\beta}' \mathbf{x}_i,$$

where g is some suitable link function and $\boldsymbol{\beta}$ is a vector of p unknown parameters. The usual (canonical) link function for the binomial distribution is the logit link

[1] A Bernoulli trial is a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted.

$$g(\mu_i) = \log\left(\frac{\mu_i}{m_i - \mu_i}\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

which corresponds to modelling on the log-odds scale with parameters corresponding to log odds-ratios. Other common choices of link function for proportion data are the probit

$$g(\mu_i) = \Phi^{-1}(\mu_i/m_i) = \Phi^{-1}(\pi_i),$$

based on an underlying normal tolerance distribution for the probability of a positive response, and the complementary log-log (CLL) link

$$g(\mu_i) = \log\{-\log(1-\mu_i/m_i)\} = \log\{-\log(1-\pi_i)\}.$$

The probit and logit links are very similar and are symmetric in π and $1-\pi$, while the CLL link is not symmetric and can lead to rather different fits in certain cases. The probit has along history in biological assays and dose-response studies, see Finney (????), although in gneral the logit is now preferred.

Model comparison and inference for generalized linear models is based on the *Analysis of deviance*, generalizing ideas from ANOVA and first introduced by Nelder and Wedderburn (1972). The residual deviance compares a fitted model to a saturated model that reproduces the observed data and for the binomial model is given by

$$D_B = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right],$$

where $\hat{\mu}_i$, $i = 1,2,\ldots,n$, is the fitted value for the model of interest. The deviance D_B is a measure of goodness-of-fit of the fitted model with p estimated parameters. A more traditional goodness-of-fit statistic is the Pearson X^2 statistic

$$X_B^2 = \sum_{i=1}^{n}\frac{m_i(y_i - \hat{\mu}_i)}{\hat{\mu}_i(m_i - \hat{\mu}_i)}.$$

Asymptotically, D_B and X_B^2 are equivalent and both have a χ^2 distribution with $n-p$ degrees of freedom. Jørgensen (2006) recommends to use X_B^2 instead of D_B as a measure of goodness-of-fit. ***Why??*** Then for a well fitted model one would expect $X_B^2 < \chi^2_{(n-p),(1-\alpha)}$ at α significance level or approximately $X_B^2 \approx n-p$.

To test the effect of a covariate in a model the analysis of deviance requires using the difference between the residual deviances for the fitted models with (D_p) and without (D_q) the covariate. This is simply the log likelihood statistic, which under the null hypothesis that the covariate is not important, i.e. the associated parameters are zero, has an asymptotic χ^2 distribution with $p-q$ degrees of freedom, where q is the number of parameters of the model without the covariate of interest. Then, if $D_q - D_p > \chi^2_{(p-q),(1-\alpha)}$ we say that there is evidence for the inclusion of the covariate of interest at α significance level.

To complete the checking of a model it is useful to use some diagnostics plots suchas:

- plot of the observed values versus fitted values;
- plot of the components of deviance versus fitted values and
- (half-)normal plot for the components of deviance with a simulation envelope (Hinde and Demétrio, 1998).

Example: The neuropteran *Chrysoperla externa* is a predator that acts as a natural enemy of the brown citrus aphid, *Toxoptera citricida*, which is among the most important citrus' pests worldwide. A substance called "lime sulfur" is a product made of calcium polysulfides and used to control fungi, bacteria and insects that live on trees. In an optimal scenario, one could use the lime sulfur and *C. externa* together to control the population size of *T. citricida* and the lime sulfur would not affect the predator as much as it would affect the pest.

Table 1 presents data from an experiment that used first-instar larvae of *Chrysoperla externa* to assess the effects of lime sulfur on the development of the predator (???). Twenty-four Orange Jessamine (*Murraya paniculata*) plants were sprayed with different concentrations of lime sulfur and six first-instar larvae were transferred to each plant. They were observed until they reached the second instar and the number of larvae that died was recorded. The experiment was set up in a completely randomized design with four treatments: distilled water (control), lime sulfur at 60ppm, lime sulfur at 600ppm and lime sulfur at 6000ppm.

Why is m_i not 6 for all i?

Table 1 *Chrysoperla externa* mortality data (table entries y_i/m_i).

Treatment	Replicates					
Control	0/2	0/5	1/4	0/5	0/3	0/6
60ppm	0/6	2/6	0/5	0/5	0/4	1/5
600ppm	1/3	1/5	2/3	2/6	1/5	0/4
6000ppm	0/1	0/5	1/5	2/4	3/7	2/5

Fitting a standard linear binomial logit model to these data, using R (R Core Team, 2013)

```
dead <- c(0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 1,
          1, 1, 2, 2, 1, 0, 0, 0, 1, 2, 3, 2)
alive <- c(2, 5, 3, 5, 3, 6, 6, 4, 5, 5, 4, 4,
           2, 4, 1, 4, 4, 4, 1, 5, 4, 2, 4, 3)
treat <- rep(c(0,60,600,6000),each=6)
resp <- cbind(dead, alive)
fit <- glm(resp ~ log(treat+1) + factor(treat), family=binomial)
anova(fit, test="Chisq")
```

we have the ANODEV presented in Table **??**.

Arbitrary handling of the 0 control dose - why is this appropriate? Natural mortatility - well no evidence here, but

Do I have these data? I think that you need a more complete table here, also X^2_B values given comments earlier.

Table 2 Analysis of deviance for the *Chrysoperla externa* mortality data.

Sources of variation	Df	Deviance	Residual Df	Residual Deviance	p-value
Linear trend	1	8.73	22	22.86	$< 0.01^*$
Lack of fit	2	0.91	20	21.95	0.63

The value of the residual deviance 21.95 on 20 df compared with $\chi^2_{20,0.95} = 31.41$ gives evidence of a well fitted model. This is confirmed by the half-normal plot present in Figure 1(a). The fitted curve with the observed proportions is presented in Figure 1(b).

need to jitter the points to see the replicates - also the plotting of the 0 dose is arbitrary so does the fitted curve really make sense?

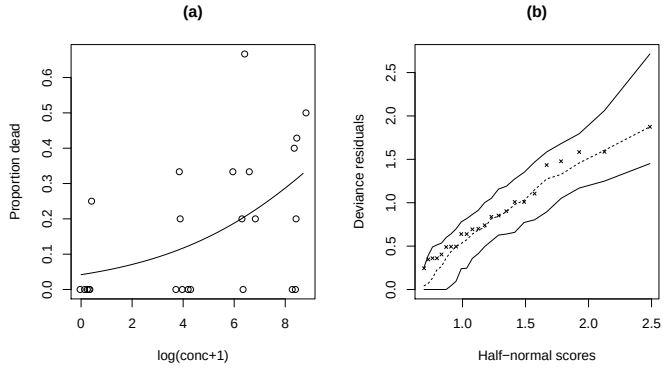

Fig. 1 *Chrysoperla externa* mortality data. (a) Half-normal plots: $\times$ – data; —- simulated envelope (b) fitted curve with observed proportions.

2.2 Overdispersion models

Once we have established that a particular dataset may exhibit overdispersion we need to think about extending our basic model to take account of this fact. As we have discussed there are many different possible causes of overdispersion and consequently a number of different models and associated estimation methods have been proposed (see Hinde and Demétrio, 1998 for a review). For binomial data, Collett (1991) gives a good practical introduction to some of these methods, following the work of Williams (1982, 1996).

2.2.1 Quasilikelihood

One of the simplest means to allow for overdispersion is to replace the mean-variance function of the original model by a more general form, typically involving additional parameters.

A constant overdispersion model replaces (2) by

$$\mathsf{Var}\,(Y_i) = \phi m_i \pi_i (1-\pi_i). \tag{3}$$

The constant overdispersion factor ϕ indicates that the overdispersion for observation Y_i depends on neither the sample size m_i nor the true response probability π_i. This is often referred to as the heterogeneity factor model, see Finney (1971).

Example: This experiment assessed the insecticide action of organic extracts of *Annona mucosa* (Annonaceae) by evaluating damage to corn grains which were infested with *Sitophilus zeamais* (Coleoptera: Curculionidae), the major pest of stored maize in Brazil (Ribeiro et al, 2013). Petri dishes containing 10g of corn were treated with extracts that were prepared with different parts of the plant (seeds, leaves and branches) at the concentration of 1500 mg/kg. To each Petri dishes 20 *Sitophilus zeamais* adults were added and, after 60 days, an assessment was made of the proportion of damaged corn grains. The experiment was set up in a complete randomized design with 10 replicates. The four treatments were leaf extract, branch extract, seed extract and distilled water (control).

2.2.2 Beta-binomial

Adopting a two-stage model, if we assume that $Y_i|P_i \sim \mathrm{Bin}(m_i, P_i)$, where the P_i's are now taken as random variables with Beta(α_i, β_i) distributions with $\alpha_i + \beta_i$ constant, then unconditionally, we have

$$\mathrm{E}(Y_i) = m_i \pi_i$$

and

$$\mathsf{Var}\,(Y_i) = m_i \pi_i (1-\pi_i)[1+\phi(m_i-1)]. \tag{4}$$

Example

2.2.3 Logistic/Probit normal

The beta-binomial model assumes that the P_i have a beta distribution. Another possibility is to assume that the linear predictor, η_i, has some continuous distribution. If this distribution is taken to be in the location-scale family then this corresponds to including an additive random effect in the linear predictor and we can write

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma z_i$$

where z_i is assumed to be from the standardized form of the distribution. Most commonly z is taken to be normally distributed leading to the logistic-normal and probit-normal models. The probit-normal has a particularly simple form as the individual binary responses can be considered as arising from a threshold model for a normally distributed latent variable, see McCulloch (1994).

It is possible to show that

$$\mathrm{E}(Y_i) = m_i\pi_i$$

and that the variance function can be approximated by

$$\mathsf{Var}\,(Y_i) \approx m_i\pi_i(1-\pi_i)[1+\sigma^2(m_i-1)\pi_i(1-\pi_i)], \tag{5}$$

which Williams (1982) refers to as a type III variance function.

Example

2.3 Zero-inflated models

For proportion data based sample sizes m_i, we can modify the basic binomial distribution to give a *zero-inflated binomial* (ZIB) distribution with

$$\mathsf{Pr}(Y_i = y_i) = \begin{cases} \omega + (1-\omega)(1-\pi_i)^{m_i} & y = 0 \\ (1-\omega)\binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i} & y_i = 1,\ldots,m_i. \end{cases} \tag{6}$$

The mean is given by $\mathrm{E}(Y_i) = m_i(1-\omega)\pi_i$ and the variance can be written as

$$\begin{aligned} \mathsf{Var}\,(Y_i) &= (1-\omega)\left\{m_i\pi_i(1-\pi_i) + \omega(m_i\pi_i)^2\right\} \\ &= m_i(1-\omega)\pi_i\left\{1-(1-\omega)\pi_i\right\} + m_i\pi_i^2\omega(1-\omega)(m_i-1) \end{aligned}$$

and since the latter term is non-negative this model is overdispersed compared to a binomial model.

Example:

3 Models for count data

One very common form of data collection is simple counting giving observed counts over a fixed area, period of time, volume, etc. Typical examples include:

- ecological diversity studies using numbers of insect species in quadrats at different locations;
- physiology experiments, assessing fecundity using number of eggs laid per female;

– behavioural studies to analyse reproduction patterns using number of matings;
– biological control assays, with the numbers of parasitised eggs or numbers of attacked prey as the response of interest.

3.1 Poisson models

For counts over time or space a very simple model is to assume that events happen independently, singly, and at random at some constant underlying rate. Considering events over time, such as radioactive emissions, ***Can't we have a more relevant illustration here?*** if we write λ for the average rate per unit time, then, under these assumptions, the distribution for the number of events, $Y(t)$, in an interval of length t is Poisson(λt) with probabilities

$$\mathsf{Pr}(Y(t) = y) = \frac{e^{-\lambda t}(\lambda t)^y}{y!}, \qquad y = 0, 1, 2, \ldots . \tag{7}$$

This result is a consequence of the Poisson approximation to the binomial distribution and is given in many elementary texts on probability such as Ross (1993). The mean and variance are e$Y(t) = \lambda t$ and $\mathsf{Var}\,(Y(t)) = \lambda t$ and are equal; this is an important practical characterisation of the Poisson distribution. A similar model applies for counts over space, with the parameter λ giving the rate per unit area or volume.

In many simple applications, counts will be observed over identical time periods, areas, etc. In this case, we can use a standard Poisson(μ) distribution for a count Y with

$$\mathsf{Pr}(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, \ldots \tag{8}$$

where μ is the rate for the region of observation and

$$\mathrm{E}(Y) = \mathsf{Var}\,(Y) = \mu.$$

glms and offsets, etc

3.2 Overdispersion models

Overdispersed Poisson data are discussed, for example, in Breslow (1984) and Lawless (1987). More general discussions of overdispersion are also to be found in McCullagh and Nelder (1989) and Lindsey (1995).

3.2.1 Quasilikelihood

As in Section 2.2.1 we suppose that the random variables Y_i represent counts with means μ_i.

3.2.2 Constant overdispersion

A constant overdispersion model replaces **??** by

$$\mathsf{Var}\,(Y_i) = \phi\mu_i. \tag{9}$$

Example

3.2.3 Negative binomial

A two-stage model assumes that $Y_i|\theta_i \sim \text{Pois}(\theta_i)$ where the θ_i's are random variables following $\Gamma(k,\lambda_i)$ distributions. Then, unconditionally we have $\text{E}(Y_i) = k/\lambda_i = \mu_i$ and

$$\mathsf{Var}\,(Y_i) = \mu_i + \mu_i^2/k. \tag{10}$$

Example

3.2.4 Poisson/normal

Proceeding as for the binomial model we can also consider including a random effect in the linear predictor. Using a Poisson log-linear model and a normally distributed random effect leads to the Poisson-normal model, see Hinde (1982) for details of maximum likelihood estimation. To obtain the variance function here, we can specify the model as

$$Y_i \sim \text{Pois}(\lambda_i) \quad \text{with} \quad \log\lambda_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma Z_i$$

where $Z_i \sim N(0,1)$, which gives unconditionally

$$\begin{aligned}\text{E}(Y_i) &= e^{\mathbf{x}_i^T\boldsymbol{\beta}+\frac{1}{2}\sigma^2} := \mu_i \\ \mathsf{Var}\,(Y_i) &= e^{\mathbf{x}_i^T\boldsymbol{\beta}+\frac{1}{2}\sigma^2} + e^{2\mathbf{x}_i^T\boldsymbol{\beta}+\sigma^2}(e^{\sigma^2}-1) = \mu_i + k'\mu_i^2.\end{aligned}$$

Example

3.3 Zero-inflated models

For count data we take the Poisson distribution as as the underlying probability model and obtain a *zero-inflated Poisson* (ZIP) distribution with

$$\mathsf{Pr}(Y=y)=\begin{cases}\omega+(1-\omega)e^{-\theta} & y=0\\ (1-\omega)\dfrac{e^{-\theta}\theta^{y}}{y!} & y=1,2,\ldots\end{cases} \tag{11}$$

The mean is given by $\mu=\mathrm{e}Y=(1-\omega)\theta$, while the variance is

$$\mathsf{Var}\,(Y)=(1-\omega)\theta(1+\omega\theta)=\mu+\mu^{2}\frac{\omega}{(1-\omega)}$$

which is greater than μ, unless $\omega=0$, and has the same quadratic form as the negative binomial variance. In this sense we may think of it as a model for overdispersed count data, but data in which the overdispersion arises in a very specific way, through an excess of zeros.

Example

4 Models for time to event data

4.1 Discrete survival models

4.1.1 Ordinal data - multinomial models

4.2 Continuous survival models

4.2.1 Parametric models

4.2.2 Cox proportional hazards model

4.2.3 Frailty

References need completing, expanding, etc — I guess that we cannot use BiBTeX because of the style file?

References

1. Battel, A.P.M.B.: Dinâmica de predação e resposta funcional em *Chrysoperla externa* (Neuroptera: Chrysopidae) sobre *Toxoptera citricida* (Hemiptera: Aphididae) aplicada à citricultura orgânica. Master's dissertation.
2. Collet, D.: Modelling binary data. Chapman and Hall, London (1994).
3. Dobson, A.J.: An Introduction to Generalized Linear Models. Chapman and Hall, London (2002).
4. Hinde, J.; Demétrio, C.G.B.: Overdispersion: models and estimation. Computational Statistics and Data Analysis, **27**: 151-170 (1998).
5. Jørgensen (2013)
6. McCullagh, P; Nelder, J.: Generalized Linear Models. (2nd ed.) Chapman and Hall, London (1989).
7. RIBEIRO, L.P.; VENDRAMIN, J.D.; BICALHO, K.U.; ANDRADE, M.S.; FERNANDES, J.B.; MORAL, R.A.; DEMÉTRIO; C.G.B. *Annona mucosa* Jacq. (Annonaceae): A promising source of bioactive compounds against *Sitophilus zeamais* Mots. (Coleoptera: Curculionidae). Journal of Stored Products Research 55:6-14, 2013.