Overdispersion:
Models and estimation

John Hinde[a,*], Clarice G.B. Demétrio[b]

[a]MSOR Department, Laver Building, University of Exeter, North Park Road,
Exeter EX4 4QE, UK
[b]DME/ESALQ, University of São Paulo, 13418-900 Piracicaba SP, Brazil

Abstract

Overdispersion models for discrete data are considered and placed in a general framework. A distinction is made between completely specified models and those with only a mean-variance specification. Different formulations for the overdispersion mechanism can lead to different variance functions which can be placed within a general family. In addition, many different estimation methods have been proposed, including maximum likelihood, moment methods, extended quasi-likelihood, pseudo-likelihood and non-parametric maximum likelihood. We explore the relationships between these methods and examine their application to a number of standard examples for count and proportion data. A simple graphical method using half-normal plots is used to examine different overdispersion models. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Generalized linear models; Overdispersion; Binomial; Beta-binomial; Poisson; Negative binomial

1. Introduction

In applying standard generalized linear models it is often found that the data exhibit greater variability than is predicted by the implicit mean-variance relationship. This phenomenon of *overdispersion* has been widely considered in the literature, particularly in relation to the binomial and Poisson distributions. Failure to take account of this overdispersion can lead to serious underestimation of standard errors and misleading inference for the regression parameters. Consequently, a number of models and associated estimation methods have been proposed for handling such data. For binomial data, Collett (1991) gives a good practical introduction to some of these

___

* Corresponding author.

methods, following the work of Williams (1982, 1996). 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).

There are many different specific models for overdispersion, which arise from alternative possible mechanisms for the underlying process. We can broadly categorise the approaches into the following two groups:

(i) Assume some more general form for the variance function, possibly including additional parameters.

(ii) Assume a two-stage model for the response. That is, assume that the basic response model parameter itself has some distribution.

Models of type (i) may not correspond to any specific probability distribution for the response but may be viewed as useful extensions of the basic model. The regression parameters can be estimated using quasi-likelihood methods with some ad hoc procedure for estimating any additional parameters in the variance function. An example of this is the use of a heterogeneity factor in overdispersed binomial data.

Type (ii) models lead to a compound probability model for the response and, in principle, all of the parameters can be estimated using full maximum likelihood. A standard example is the use of the negative binomial distribution for overdispersed count data. However, in general, the resulting compound distribution takes no simple form and approximate methods of estimation are often used. A variant of the second approach, that removes the need to make any specific assumptions about the second stage distribution, is to use non-parametric maximum likelihood (NPML) estimation of the compounding distribution, as advocated by Aitkin (1995, 1996).

In Section 2 we consider models for overdispersed binary data and introduce the various methods of estimation. The use of these is illustrated on two standard examples. In Section 3 we introduce models for overdispersed count data. More complex models for overdispersion are discussed in Section 4. Finally, in Section 5 we use a simple graphical procedure for diagnostic assessment of overdispersed models.

2. Binary data

2.1. Models and estimation

In studying a binary response suppose that the random variables Y_i represent counts of successes out of samples of size m_i, $i = 1, \ldots, n$. If we write $E[Y_i] = \mu_i = m_i \pi_i$, then a generalized linear model allows us to model the expected proportions π_i in terms of explanatory variables x_i through

$$g(\pi_i) = \boldsymbol{\beta}' x_i,$$

where g is some suitable link function and $\boldsymbol{\beta}$ is a vector of p unknown parameters. The usual error specification is $Y_i \sim \text{Bin}(m_i, \pi_i)$ with variance function

$$\text{Var}(Y_i) = m_i \pi_i (1 - \pi_i). \tag{1}$$

However, when overdispersion is present the variance will be greater than this. In the following sections we will consider various generalizations of this basic variance function in terms of the two approaches outlined in Section 1.

2.1.1. Constant overdispersion

The standard quasi-likelihood approach, which requires only the specification of the first two moments, replaces (1) by

$$\text{Var}(Y_i) = \phi m_i \pi_i (1 - \pi_i). \tag{2}$$

The constant overdispersion factor ϕ is estimated by equating the Pearson X^2 statistic from a binomial fit to its degrees of freedom (McCullagh and Nelder, 1989). The estimates for the linear predictor parameters are the same as those for the binomial, although the standard errors will be inflated by the overdispersion factor. This is the heterogeneity factor model, see Finney (1971).

2.1.2. Beta-binomial variance function (Williams type II)

Adopting a two-stage model, if we assume that $Y_i \sim \text{Bin}(m_i, P_i)$, where the P_i's are now taken as random variables with $E(P_i) = \pi_i$ and $\text{Var}(P_i) = \phi \pi_i (1 - \pi_i)$ then, unconditionally, we have

$$E(Y_i) = m_i \pi_i$$

and

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

A special case of this is the beta-binomial distribution, which is obtained by assuming that the P_i's have $\text{Beta}(\alpha_i, \beta_i)$ distributions with $\alpha_i + \beta_i$ constant. For the beta-binomial distribution full maximum likelihood estimation is possible, see Crowder (1978). Instead of using the full form of the beta-binomial likelihood it is also possible to use estimation methods based on just the first two moments. This removes the problem of specifying a particular distribution for the P_i's.

Nelder and Pregibon (1987) consider an extension of the quasi-likelihood model in which the variance is now specified as $\text{Var}(Y_i) = \phi_i V(\mu_i)$, where both ϕ_i and the variance function $V(\cdot)$ may depend upon additional parameters. For observed responses y_i, $i = 1, \ldots, n$, to estimate the unknown parameters in the mean and variance models they propose maximising the extended quasi-likelihood (EQL) function

$$Q^+ = -\frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{D(y_i, \mu_i)}{\phi_i} + \log(2\pi\phi_i V(y_i)) \right\},$$

where D is the deviance function

$$D(y, \mu) = -2 \int_{y}^{\mu} \frac{(y - t)}{V(t)} \, dt.$$

For the beta-binomial model we can take the variance function $V(\mu_i)$ as $\mu_i(1 - \mu_i/m_i)$, the usual binomial variance function, and $\phi_i = 1 + \phi(m_i - 1)$ giving a deviance

function, D, identical to that for the binomial model. Maximizing Q^+ over the regression parameters β gives a set of weighted binomial estimating equations with weights $1/\phi_i = 1/[1 + \phi(m_i - 1)]$. These are the quasi-likelihood equations for a known value of the overdispersion parameter ϕ. For the estimation of ϕ we need to solve

$$\sum_{i=1}^{n} \left\{ \frac{D(y_i, \mu_i)}{\phi_i} - 1 \right\} \frac{\mathrm{d}\log(\phi_i)}{\mathrm{d}\phi} = \sum_{i=1}^{n} \left\{ \frac{D(y_i, \mu_i) - \phi_i}{\phi_i^2} \right\} \frac{\mathrm{d}\phi_i}{\mathrm{d}\phi} = 0.$$

The second form of this equation shows that we can obtain an estimate for ϕ by fitting a gamma model using the deviance components as the y-variable, an identity link and taking the linear model to have a fixed intercept (offset) of 1 and $m_i - 1$ as the explanatory variable. An approximate standard error is obtained for ϕ by setting the scale to 2, corresponding to modelling χ_1^2 responses (McCullagh and Nelder, 1989). We iterate between these two sets of estimating equations for β and ϕ until convergence, giving parameter estimates and standard errors, which are correct because of the asymptotic independence of $\widehat{\beta}$ and $\widehat{\phi}$.

Brooks (1984) suggests a mixed strategy in which quasi-likelihood estimation for β is combined with maximum likelihood estimation for ϕ. This involves replacing the above estimating equation for ϕ by the beta-binomial likelihood score equation, with β set equal to its current estimate. This equation is easily solved using a Newton–Raphson-type iteration, although it may be necessary to reduce the step length to avoid divergence or values outside of the feasible region for ϕ. In practice, this approach seems to give estimates for (β, ϕ) which are very close to the full maximum likelihood estimates. They will not be exact maximum likelihood estimates, as for fixed ϕ the beta-binomial distribution is not in the exponential family and so the quasi-likelihood estimates for β with the beta-binomial variance function do not exactly maximize the beta-binomial likelihood.

An alternative to extended quasi-likelihood is the pseudo-likelihood (PL) approach of Carroll and Ruppert (1988). Here estimates of β are obtained by generalized least squares, which if iterated is equivalent to quasi-likelihood estimation for given values of ϕ_i. The estimation of additional parameters in the variance is based on the maximization of

$$P = -\frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{(y_i - \mu_i)^2}{\phi_i V(\mu_i)} + \log\left(2\pi\phi_i V(\mu_i)\right) \right\}.$$

This is of the same form as Q^+ but with the deviance increments replaced by the squared Pearson residuals and $V(y_i)$ by $V(\mu_i)$; it corresponds to a normal likelihood function for the residuals. For the beta-binomial distribution the estimating equation for ϕ is

$$\sum_{i=1}^{n} \left\{ \frac{(y_i - \mu_i)^2}{\phi_i V(\mu_i)} - 1 \right\} \frac{\mathrm{d}\log(\phi_i)}{\mathrm{d}\phi} = \sum_{i=1}^{n} \left\{ \frac{r_i^2 - \phi_i}{\phi_i^2} \right\} \frac{\mathrm{d}\phi_i}{\mathrm{d}\phi} = 0,$$

where $r_i = (y_i - \mu_i)/\sqrt{V(\mu_i)}$, the unscaled generalized Pearson residuals. This equation can be solved in the same way as the EQL estimating equation by fitting a gamma model to the squared Pearson residuals.

Another possibility, discussed by Moore (1986), is to use a simple moment method, replacing the pseudo-likelihood estimating equation for ϕ by the following unbiased estimating equation

$$\sum_{i=1}^{n} \left\{ \frac{(y_i - \mu_i)^2}{\phi_i V(\mu_i)} - 1 \right\} = 0.$$

This moment method corresponds to solving $X^2 = n$, where X^2 is the generalized Pearson χ^2-statistic. A variant of this is to solve $X^2 = n - p$, where $p = \dim(\beta)$, which amounts to a degrees of freedom correction for the parameters in the regression model for the mean. This equation can be solved iteratively using either Newton–Raphson or fixed-point type methods. In an early paper on overdispersed binomial models, Williams (1982) proposes estimating ϕ by solving $X^2 = E[X^2]$, which gives simple one-step update for ϕ. On iterating this with quasi-likelihood estimation for β we obtain the same estimates as the degrees of freedom corrected moment method, since on convergence $E[X^2] = n - p$.

For the case in which all of the sample sizes are equal to m, the beta-binomial variance function (3) reduces to constant overdispersion, as in (2). The weights in the quasi-likelihood estimating equations for β are all constant and so these equations reduce to those for the standard binomial model. The estimation of ϕ is also greatly simplified as $d \log(\phi_i)/d\phi = (m - 1)/[1 + \phi(m - 1)]$ is now constant. Thus, EQL reduces to equating the binomial model scaled deviance to n, while PL uses the Pearson X^2 as in Moore's method. In the case of unequal sample sizes it can be seen that the moment method is essentially a weighted version of the PL estimating equation. This suggests that a similar approach may also be taken in EQL, equating the scaled deviance to n, although the resulting equation is not unbiased as $E[D(y_i, \mu_i)] \neq \phi_i$. To take account of the estimation of β it is possible to apply a degrees of freedom correction to EQL and PL by including the factor $(n - p)/n$ before the second term in the expression for Q^+ or P, see McCullagh and Nelder (1989, p. 362).

A conceptually different model for overdispersion is to assume that the individual binary responses are not independent. Assuming a constant correlation ρ leads to a model for the Y_i with

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

which is of exactly the same form as (3). However, it is now possible for ρ to be negative $(-1/(m_{(n)} - 1) < \rho < 1)$ corresponding to underdispersion. Using the mean-variance specification this model is fitted precisely as above. It is also possible to extend the beta-binomial distribution to handle underdispersion (Prentice, 1986).

2.1.3. Logistic-normal (Williams type III) and related models

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

$$\eta_i = \beta' x_i + \sigma z_i,$$

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). A general approach to the estimation of models of this type is to use the EM-algorithm with Gaussian quadrature to integrate over the normal distribution, following the same approach given by Hinde (1982) for the Poisson distribution. Williams (1982) shows that, for small values of σ, the logistic-normal model can be approximated by a quasi-likelihood model with a variance function of the form

$$\mathrm{Var}(Y_i) = m_i\pi_i(1 - \pi_i)[1 + \sigma^2(m_i - 1)\pi_i(1 - \pi_i)],\tag{4}$$

which he refers to as type III. EQL, PL and moment methods can be used for the estimation of σ^2 taking $\phi_i = 1 + \sigma^2(m_i - 1)\pi_i(1 - \pi_i)$.

Aitkin (1995, 1996) replaces the assumed distributional form for z by a discrete mixing distribution. The computation again involves an EM-type algorithm in which the mixing distribution is assumed to be a discrete distribution on a specified number of points and the weights and points for the quadrature become additional parameters in the model. This results in a non-parametric maximum likelihood (NPML) estimate of this distribution together with the regression parameter estimates.

2.1.4. A general variance function

The various variance functions considered above can be seen to be special cases of the following general form

$$\mathrm{Var}(Y_i) = m_i\pi_i(1 - \pi_i)\left[1 + \phi(m_i - 1)^{\delta_1}\{\pi_i(1 - \pi_i)\}^{\delta_2}\right].\tag{5}$$

The standard binomial model corresponds to $\phi = 0$; $\delta_1 = \delta_2 = 0$ gives the constant overdispersion model; $\delta_1 = 1$, $\delta_2 = 0$, the beta-binomial variance function, and $\delta_1 = \delta_2 = 1$, the Williams type III model. A set of GLIM4 macros (Hinde, 1996) allows the fitting of overdispersed binomial data with a variance function $\mathrm{Var}(Y_i) = m_i\pi_i(1 - \pi_i)[1 + \phi f(m_i, \pi_i)]$, where f is a function specified by the user. Hence, the above general form can be used for specified values of δ_1 and δ_2.

2.2. Examples

Here we will consider the application of the various methods described above to several standard examples from the overdispersion literature. This extends the comparison by Liang and McCullagh (1993), who only consider constant overdispersion and the beta-binomial-type variance function.

2.2.1. Orobanche germination data

Crowder (1978) presents data from a study of the germination of two species of *Orobanche* seeds grown on 1/125 dilutions of two different root extract media (cucumber or bean) in a 2×2 factorial layout with replicates, see also Collett (1991).

Table 1
Orobanche data: deviances with overdispersion estimated from maximal model

		Binomial	Constant	Beta-binomial		
Source	d.f.	ML	QL	ML	EQL	PL
Extract | Species	1	56.49	30.34	32.69	31.68	31.37
Species | Extract	1	3.06	1.64	2.88	2.84	2.85
Species.Extract	1	6.41	3.44	4.45	4.40	4.62
$\hat{\phi}$			1.862	0.012	0.013	0.013

		Beta-binomial			Logistic-normal	
Source	d.f.	Moment[a]	EQL[a]	PL[a]	Moment[a]	ML
Extract | Species	1	22.95	24.67	24.76	21.49	31.33
Species | Extract	1	2.64	2.69	2.72	2.54	2.85
Species.Extract	1	3.54	3.72	3.98	3.52	4.44
$\hat{\phi}$		0.025	0.022	0.021	—	—
$\hat{\sigma}^2$		—	—	—	0.108	0.056

[a](df corrected).

The data consist of the number of seeds and the number germinating for each replicate. Interest focusses on the possible differences in germination rates for the two types of seed and root extract and whether there is any interaction. Table 1 presents results for different models and estimation methods with the overdispersion parameter estimated from the interaction model and fixed for all sub-models. An alternative strategy is to re-estimate the dispersion parameter for each model; this is considered later. Note that the overdispersion parameter estimates are not all comparable as they relate to different overdispersion models.

The residual deviance for the interaction model using the standard binomial fit is 33.28 on 17 df, giving strong evidence of overdispersion. Using a constant overdispersion model and quasi-likelihood gives a non-significant interaction term (deviance = 3.44), with extract as the only important factor (deviance = 30.34). The conclusions are less clear cut for the other overdispersion models with the interaction being marginally significant. For the beta-binomial variance function the only real differences between the estimation methods are whether we use the degrees of freedom correction or not; without any correction the EQL and PL results are very similar to the maximum likelihood fit (ML) and, indeed, using an uncorrected moment method also leads to very similar estimates. The slight difference between the degrees of freedom corrected moment and PL methods is due to the different forms of weighting in the estimating equation for ϕ. The sample sizes here vary from 4 to 81 giving weights which vary by a factor of 10 for a typical value of ϕ, although the impact on the final estimate is slight. Interestingly, if we parameterize the constant overdispersion model as $1 + \gamma(\bar{m} - 1)$, where the mean sample size is $\bar{m} = 39.6$, we obtain $\hat{\gamma} = 0.022$, which is very close to the degrees of freedom corrected estimates for the beta-binomial variance function. The differences between

the maximum likelihood and moment method fits for the logistic-normal can also be attributed to the degrees of freedom correction. Notice that, as the fitted proportions for the `Species*Extract` model are not extreme (from 0.36 to 0.68 with overall proportion 0.51) the results from using a logistic-normal variance function (4) are very similar to those using a beta-binomial form (3) – for all fitted proportions $\hat{\pi}_i$, the logistic-normal moment estimates give $\hat{\sigma}^2 \hat{\pi}_i (1 - \hat{\pi}_i) \approx 0.025$, the estimate of ϕ under the beta-binomial model.

Comparing the beta-binomial and binomial models the change in deviance for the additional parameter is 2.34 on 1 degree of freedom giving no evidence for the beta-binomial overdispersion function. Nelder and Pregibon (1987) make a similar observation in considering the EQL fit. Liang and McCullagh (1993) conduct a formal test between the constant overdispersion and beta-binomial overdispersion models and are unable to choose between them. The same conclusion results from using the general form of the variance function introduced above and looking at the profile likelihood for δ_1 with $\delta_2 = 0$.

There is considerable confusion about residual deviances for overdispersion models. In general, these provide no information about the fit of the model, because of the estimation of the overdispersion parameters. Deviances for the beta-binomial and logistic-normal models are often given with respect to a binomial saturated model and, while this is a useful device for comparisons with the standard binomial model, it does not provide a goodness of fit measure. For the beta-binomial family it is possible to calculate a true deviance for a *fixed* value of the overdispersion parameter, but when the parameter is estimated, not surprisingly, this always seems to result in values close to the degrees of freedom.

Using the NPML approach the mixing distribution is considered as a nuisance parameter and estimated for each model. Table 2 compares the deviance differences from this model with results from other overdispersion modelling methods when the overdispersion parameter is re-estimated for each submodel. The results are all similar. For the interaction model the NPML fit gives a two-point mixing distribution with a variance of 0.08, comparable to the variance estimate of 0.056 for the logistic-normal fit. A plot of the component probabilities against the binomial model residuals shows a strong monotonic relationship, indicating that the mixing distribution is picking up overdispersion.

It is interesting to note that the results from Tables 1 and 2, for using a fixed overdispersion parameter or re-estimating it in each model, are fairly similar. Although, as would be expected, all of the deviance contributions are reduced in Table 2

Table 2
Orobanche data: deviances with overdispersion re-estimated for each model

Source	NPML	Logistic normal	Beta-binomial		
			ML	EQL	PL
Extract | Species	12.92	15.26	15.44	15.08	15.78
Species | Extract	1.98	2.70	2.73	2.70	2.73
Species.Extract	4.22	4.15	4.13	4.10	4.34

Table 3
Trout egg data

Location	Survival period (weeks)			
in stream	4	7	8	11
1	89/94	94/98	77/86	141/155
2	106/108	91/106	87/96	104/122
3	119/123	100/130	88/119	91/125
4	104/104	80/97	67/99	111/132
5	49/93	11/113	18/88	0/138

Table 4
Trout egg data: deviances with overdispersion estimated from maximal model

Source	d.f.	Binomial		Constant	Beta-binomial	
		Logit	CLL	CLL	Logit	CLL
Loc | Time	4	849.1	853.8	184.1	158.6	178.6
Time | Loc	3	164.1	168.8	36.4	31.4	35.9
Residual	12	64.5	59.8			
$\hat{\phi}$				4.64	0.038	0.033

as the overdispersion parameter estimate increases to account for the extra-variation of the omitted term. The fixed strategy seems more sensible and is analogous with the usual procedure for the normal model. Our primary interest will usually be in the fixed effects model and if possible we would like to obtain an estimate of pure overdispersion - this is available from the maximal model.

2.2.2. Trout egg data

Manly (1978) considers the analysis of data on the survival of trout eggs. Boxes of eggs were buried at five different locations in a stream and at four different times a box from each location was sampled and the number of surviving eggs were counted. The data are presented in Table 3 as proportions s/n, where s denotes the number of survivors and n the number of eggs in the box.

In his original analysis of this data, Manly used a log–log link function for the probability of surviving, which we reproduce here by using a complementary log–log link (CLL) for the probability of death. In Table 4 we present deviance contributions for the two factors in several models. The residual deviance for the binomial logit model is 64.5 on 12 df, while that for the CLL model is 59.8 indicating a slight preference for this link. The results for the beta-binomial variance function are based on moment estimation.

The large amount of overdispersion present here has a clear impact on the effect deviances, although both effects remain significant in all of the models. For the beta-binomial variance function with complementary log–log link we have a slightly smaller estimate of the overdispersion parameter and correspondingly larger deviance

differences for the effects. We will return to the comparison of these models in Section 5. The other point to note here is the great similarity for the complementary log–log link of the constant overdispersion and beta-binomial deviances. This is because the sample sizes here are relatively large and not too different (ranging from 86 to 155 with mean 111). If we parameterize the constant overdispersion model as $1 + \gamma(\bar{m} - 1)$, where $\bar{m}$ is the mean sample size, we obtain $\hat{\gamma}=0.033$, exactly the same as in the beta-binomial model. Because of the large and relatively similar sample sizes the other estimation approaches for the beta-binomial variance function give similar results.

The NPML approach gives a less satisfactory answer here with components picking off the observations with extreme observed proportions.

3. Overdispersion models for count data

3.1. Models and estimation

We now assume that the random variables Y_i, $i = 1,\ldots,n$, represent counts with means μ_i. The standard Poisson model assumes that $Y_i \sim \text{Pois}(\mu_i)$ with variance function

$$\text{Var}(Y_i) = \mu_i. \tag{6}$$

When there is overdispersion we need to consider variance functions which predict greater variability. A simple constant overdispersion model replaces (6) by

$$\text{Var}(Y_i) = \phi\mu_i \tag{7}$$

and estimation proceeds by quasi-likelihood as for the binomial.

3.1.1. Negative binomial type variance

A two-stage model assumes that $Y_i \sim \text{Pois}(\theta_i)$ where the θ_i's are random variables with $E(\theta_i) = \mu_i$ and $\text{Var}(\theta_i) = \sigma_i^2$. Unconditionally, we have $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \sigma_i^2$ giving an overdispersed model. Further, if the θ_i's are assumed to have a constant coefficient of variation, σ^2, we have $\text{Var}(Y_i) = \mu_i + \sigma^2\mu_i^2$, a particular form of quadratic variance function. For a fully specified model, a common assumption is that the θ_i follow a $\Gamma(k,\lambda_i)$ distribution which leads to a negative binomial distribution for the Y_i with $E(Y_i) = k/\lambda_i = \mu_i$ and

$$\text{Var}(Y_i) = \mu_i + \mu_i^2/k. \tag{8}$$

Maximum likelihood estimation for the negative binomial distribution is relatively straightforward as for fixed values of k the distribution is in the linear exponential family and estimates for regression parameters β can be obtained using the standard iteratively re-weighted least-squares (IRLS) algorithm for generalized linear models. To estimate k we can use Newton–Raphson for the score equation and cycling between the estimation of β and k we obtain the joint maximum likelihood estimates.

The asymptotic independence of $\hat{\beta}$ and $\hat{k}$ means that the standard errors for β from the IRLS fit are correct, see Lawless (1987) for details.

Using EQL for the negative binomial variance function presents some ambiguity due to different factorizations of $\mathrm{Var}(Y_i) = \phi_i V(\mu_i)$. Three obvious possibilities are

(i) $\phi_i \equiv 1$, $V(\mu_i) = \mu_i + \mu_i^2/k$;
(ii) $\phi_i = 1 + \mu_i/k$, $V(\mu_i) = \mu_i$;
(iii) $\phi_i = \mu_i + \mu_i^2/k$, $V(\mu_i) \equiv 1$.

In principle, all of these lead to different estimating equations for β, defining different iterative schemes. On convergence these all give the same estimates and the sensible approach is to use quasi-likelihood with the negative binomial variance function. For the estimation of k things are not so simple and the different formulations will lead to different estimates. Using (i) leads to an estimating equation for k similar in form to the negative binomial score equation. In cases (ii) and (iii) the parameter k appears in the scale parameter and gamma estimating equations are obtained. In (ii) Poisson deviances are used as the y-variable, while in (iii) we use Poisson Pearson residuals and this corresponds to pseudo-likelihood with estimating equation

$$\sum_{i=1}^{n} \left\{ \frac{(y_i - \mu_i)^2}{\mu_i(1 + \mu_i/k)} - 1 \right\} \frac{\mathrm{d}\log(1 + \mu_i/k)}{\mathrm{d}k} = 0.$$

Here the simple moment method gives the unbiased estimating equation

$$\sum_{i=1}^{n} \left\{ \frac{(y_i - \mu_i)^2}{\mu_i(1 + \mu_i/k)} - 1 \right\} = 0.$$

This is the form used by Breslow (1984) although he incorporated a degrees of freedom correction. This equation is easily solved for k using a fixed-point method or Newton–Raphson. Breslow uses this together with weighted Poisson regression for the estimation of β with weights $1/(1 + \tilde{\mu}_i/\tilde{k})$, where $\tilde{\mu}_i$ and $\tilde{k}$ are obtained from the previous iteration. Use of the correct negative binomial variance function is more efficient. Note again the link between the pseudo-likelihood and moment methods; if the weights in the pseudo likelihood, $\mathrm{d}\log(1 + \mu_i/k)/\mathrm{d}k = -(\mu_i/k^2)/(1 + \mu_i/k)$, are approximately constant the estimating equations will be the same. This will be true if k is small, corresponding to a large degree of overdispersion, or if all of the means μ_i are large.

Note that by assuming a different form for the gamma mixing distribution we can obtain different overdispersed Poisson models. For example, taking a $\Gamma(r_i, \lambda)$ distribution leads to $\mathrm{Var}(Y_i) = \mu_i + \mu_i/\lambda \equiv \phi\mu_i$, the constant overdispersion model. However, the resulting distribution for Y_i is not in the exponential family and so the quasi-likelihood estimates are not maximum likelihood for this distribution (Nelder and Lee (1992) give details of maximum likelihood estimation). In the same way, for the binomial distribution, by assuming that different functions of the beta distribution

parameters, α_i and β_i, are fixed, we obtain different compound distributions with different variance functions.

3.1.2. Poisson-normal and related models

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. The variance function for this model is of the form $\mathrm{Var}(Y_i) = \mu_i + k'\mu_i^2$, that is, the same as for the negative binomial distribution. In fact, with a log-link function and an additive random effect in the linear predictor, we always obtain a variance function of approximately this form for a random effect in the linear predictor, see Nelder (1985). Hence, approximate quasi-likelihood estimates are those for the negative binomial distribution. Alternatively, by using Aitkin's NPML method we can avoid any specific distributional assumption for the random effect.

3.1.3. A general variance function

A general variance function which encompasses the various models considered here is

$$\mathrm{Var}(Y_i) = \mu_i \left\{1 + \phi\mu_i^\delta\right\}, \tag{9}$$

although other natural extensions would be to consider a general quadratic variance function or a simple power function. This general formulation can be used to provide profile likelihoods for the additional parameters ϕ and δ which, in principle, provide some means of comparing the different variance functions.

3.2. Examples

3.2.1. Pump failure data

Gaver and O'Muircheartaigh (1987) present data on the numbers of failures and the period of operation, t_i (measured in 1000 s of hours) for 10 pumps from a nuclear plant. The pumps were operated in two different modes; four being run continuously and the others kept on standby and only run intermittently. To model the failure rates we need to allow for the different periods of operation. Using a log-linear model for the numbers of failures we include $\log t_i$ as an offset in the linear predictor. The Poisson model allowing for the two modes of operation has residual deviance of 71.4 on 8 df, showing a very large degree of overdispersion. Table 5 shows the results for a constant overdispersion model and a negative binomial variance function estimated by maximum likelihood (ML), EQL (using form (ii)) and PL. For the negative binomial variance function all three estimation methods give similar results. A likelihood ratio test for overdispersion comparing the negative binomial and Poisson likelihoods ($k = \infty$) has a value of 45.25 on 1 df. Since the null hypothesis of a Poisson model corresponds to a parameter value on the boundary of the parameter space, the appropriate asymptotic distribution for

Table 5
Pump data: deviances and parameter estimates

		Poisson	Constant	Negative binomial		
		ML	QL	ML	EQL	PL
Source	d.f.					
			Deviances			
Mode	1	53.1	4.8	6.1	7.3	9.4
			Parameter estimates			
mode		1.88	1.88	1.67	1.68	1.68
S.E.		0.23	0.77	0.63	0.60	0.61
$\hat{\phi}$		0.0	11.2	—	—	—
$\hat{k}$		—	—	1.30	1.46	1.39
S.E.				0.63	0.56	0.56

this statistic under the null hypothesis has a probability mass of $\frac{1}{2}$ at 0 and a $\frac{1}{2}\chi^2_{(1)}$ distribution above 0, see Lawless (1987). Here there is clearly overwhelming evidence against the Poisson assumption. A comparison of the two overdispersion models gives no clear preference, but with such a small data set that is hardly surprising.

3.2.2. Fabric fault data

These data, listed in Hinde (1982), are counts of the number of faults in rolls of fabric of different lengths. Fig. 1 shows the raw data and indicates that the mean and variance of the number of faults increase with the length of the roll. This suggests a Poisson log-linear model with log of the roll length (log l) as explanatory variable. However, this model has a residual deviance of 64.5 on 30 df, indicating overdispersion. Table 6 shows the results of fitting several overdispersion models to these data; here the overdispersion parameter is estimated in each model to allow a comparison with the NPML approach, although the estimates given in the table are just for the full model. The NPML models have estimates for the mixing distribution on just two points and again there is a strong relationship between the component probabilities and the residuals from the ordinary Poisson fit, indicating that the mixing distribution is picking up overdispersion.

The parameter estimates for the explanatory variable log l are all very similar. This is also true of the standard errors for the overdispersion models, which, as would be expected, are larger than those for the Poisson model. The deviance differences for log l are also all very similar for the overdispersion models. Some clarification is needed with regard to the residual deviances. The quasi-likelihood deviance is close to the residual degrees of freedom (30) as it must be since $\hat{\phi}$ is estimated from the residual Pearson X^2. The residual deviance for the negative binomial fit is for a specified value of k; it indicates an adequate model, but again k has been estimated. The compound Poisson model residual deviances are to a Poisson baseline; this allows a direct test with the Poisson model, but does not provide a measure of fit.
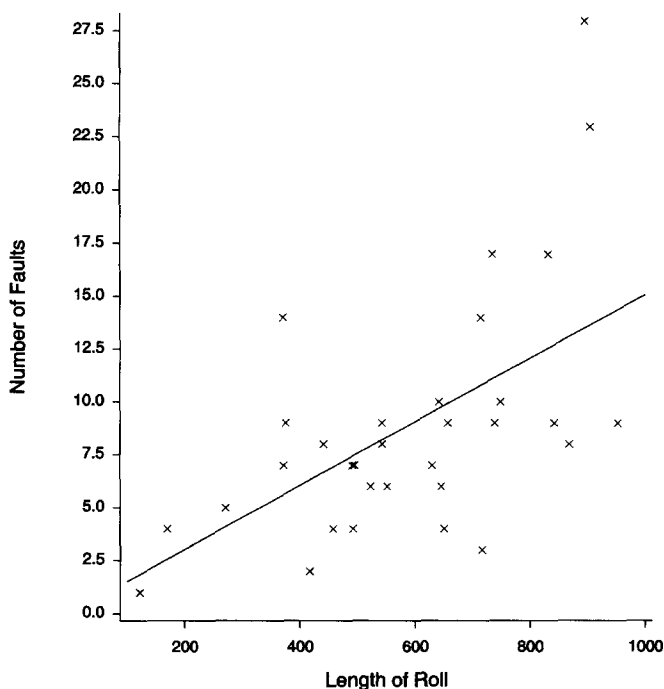
Fig. 1. Fabric fault data. × – data; — Poisson fit.

Table 6
Fabric fault data: deviances and parameter estimates

Source	Poisson ML	Constant QL	Negative binomial	Poisson normal	Poisson NPML
			Deviances		
log l	39.2	17.3	15.7	14.9	15.8
Residual	64.5	28.5	30.7	51.7	49.4
			Parameter estimates		
log l	0.997	0.997	0.938	0.922	0.800
s.e.	0.176	0.265	0.228	0.221	0.201
$\hat{\phi}$	0.0	2.27	—	—	—
$\hat{k}$	—	—	8.67	—	—
$\hat{\sigma}$	—	—	—	0.34	0.31
s.e.			0.63	0.07	

4. Extended overdispersion models

In many applications the overdispersion mechanism is assumed to be the same for all of the observations. However, in some applications it is quite conceivable that the overdispersion may be different in different subgroups of the data. Explicit models

for the variance, and hence overdispersion, are easily handled by an additional model for the scale parameter of the form

$$h(\phi_i) = \gamma' z_i$$

for some suitable link function h, usually the identity or the log. The vector of explanatory variables z_i may include covariates in the mean model giving great flexibility for joint modelling of the mean and dispersion. Estimation can proceed by either EQL or PL using a gamma estimating equation for γ as in Section 2.1.2; see McCullagh and Nelder (1989, Ch. 12), for a detailed discussion of this. A related approach is the double exponential family of Efron (1986), in which standard one-parameter exponential family distributions are extended by the inclusion of an additional parameter θ, which varies the dispersion of the family by altering the effective sample size. The dispersion parameters θ_i can again be modelled by explanatory variables and the estimation procedure is very similar to that for EQL or PL. For a simple example of modelling within this family see Fitzmaurice (1997). A very general framework for these extended models is given by the exponential dispersion models of Jorgensen (1987, 1997).

This approach to dispersion modelling is in the spirit of the our first category for handling overdispersion in Section 1. A natural way to extend the second category of models is through the addition of more complex random effects structures in the linear predictor, taking

$$\eta_i = \beta' x_i + \gamma' z_i,$$

where β is a vector of fixed effects, γ is a vector of random effects and x_i and z_i are corresponding vectors of explanatory variables. Assuming that these random effects are normally distributed gives a direct generalization of the standard linear mixed model for normally distributed responses to what is commonly called the generalized linear mixed model (GLMM). Estimation within this family is non-trivial and a number of different approaches have been proposed, including penalised quasi-likelihood (Breslow and Clayton, 1993), restricted maximum likelihood (Engel and Keen, 1994) and Bayesian methods using Markov chain Monte Carlo (Clayton, 1996). In some simple models with nested random effects, maximum likelihood estimation is possible (Anderson and Hinde, 1988) and Aitkin and Francis (1995) describe GLIM4 macros for fitting such models. In many situations the assumption of normality for the random effects is neither natural nor computationally convenient and Lee and Nelder (1996) propose an extension of GLMMs to hierarchical generalized linear models. Here, the random components can come from an arbitrary distribution, although they particularly favour the use of a distribution conjugate to that of the response. Estimation is based on h-likelihood, a generalization of the restricted maximum likelihood method used for standard normal linear mixed model. Such models are also easily handled within the Bayesian paradigm using Markov chain Monte-Carlo methods (Clayton, 1996). The non-parametric maximum likelihood approach can also be extended to these more complex models, see Aitkin (1996) and Aitkin and Francis (1995).

Table 7
Rat survival data: models with common and distinct overdispersion parameters – deviances and parameter estimates.

		Common overdispersion				
Source	d.f.	Binomial ML	Constant QL	Beta-binomial moment	Logistic-normal moment	Beta-binomial ML
Treatment	1	9.02	3.35	3.86	5.68	5.77
$\hat{\phi}$		0.0	2.69	0.20	—	0.02, 0.31
$\hat{\sigma}^2$		—	—	—	1.29	—

4.1. Examples

4.1.1. Rat survival data

Weil (1970) presents data on a toxicological study with a treatment and control group each comprising of 16 litters. The mothers in the treatment group received a diet containing the chemical of interest and the response is the number of rat pups in the litters surviving after 21 days as a fraction of the number alive at 4 days. Fitting a binomial logit model with a treatment effect results in a residual deviance of 86.2 on 30 degrees of freedom with clear evidence of overdispersion. The deviance for the treatment effect is shown in Table 7 for several overdispersion models. With both constant overdispersion and a beta-binomial variance function the treatment effect is not significant.

Examination of the data shows that there are different degrees of overdispersion in the two groups and fitting a beta-binomial model with different parameters for the two groups shows that the overdispersion parameters (0.02 and 0.31 for the control and treatment groups, respectively) differ by more than a factor of 10, however, the standard errors are very large. Allowing for this difference in variability, the treatment effect becomes significant. A similar conclusion is obtained using a logistic-normal variance function (4), where the overdispersion factor models the variance in the two groups. Liang and McCullagh (1993) note that, allowing the overdispersion to be different for the two groups, it is again not possible to choose between the constant and beta-binomial type of overdispersion. In terms of our general form of overdispersed variance function (5) it seems not to matter what value is taken for δ_1, however, taking $\delta_2 = 1$ provides a simple method of allowing for the different overdispersions in the two groups.

4.1.2. Orobanche germination data

Using EQL or PL we can fit the full interaction model Species*Extract with different overdispersion parameters for each of the four species/extract combinations. The results from both estimation methods are very similar and while the estimates for the ϕ-parameters range from 0.002 to 0.018 the standard errors are very large. The change in extended quasi-deviance is only 0.2, showing no evidence against the common overdispersion parameter model.
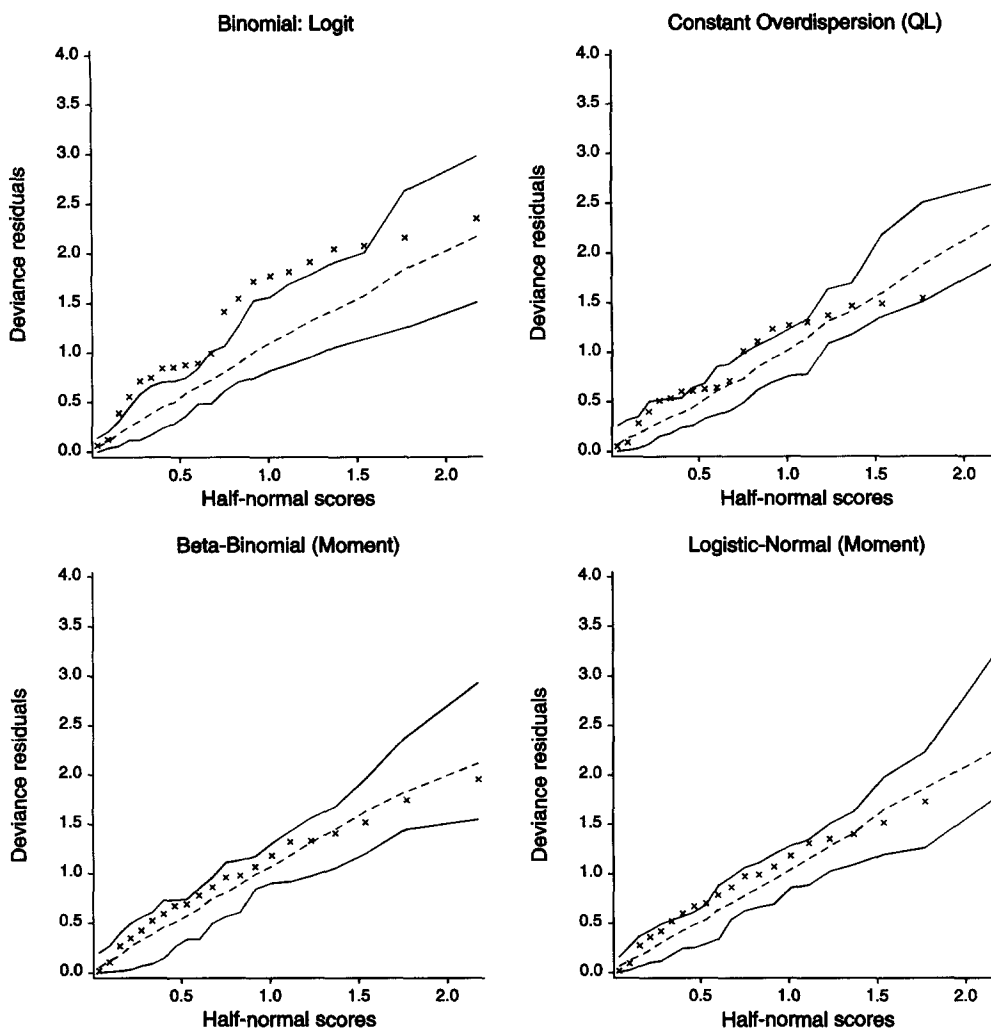
Fig. 2. Orobanche data. Half-normal plots: × – data; — simulated envelope.

5. Diagnostics

We have already mentioned problems with assessing the fit of overdispersed models, in that when an overdispersion parameter is estimated, the residual deviance or Pearson X^2 are typically close to the degrees of freedom. Similarly, residuals based on either of these quantities will be scaled and not particularly large. However, this does not mean that the residuals are no longer useful for model diagnostics, it is just that any useful information is contained in their pattern and not their absolute value. Standard residual plots can be used to explore the adequacy of the linear predictor and link function and identify outliers. A plot against the fitted values will provide an informal check of the specification of the variance function $V(\mu)$, however, this may not be helpful in choosing between overdispersion models involving the scale
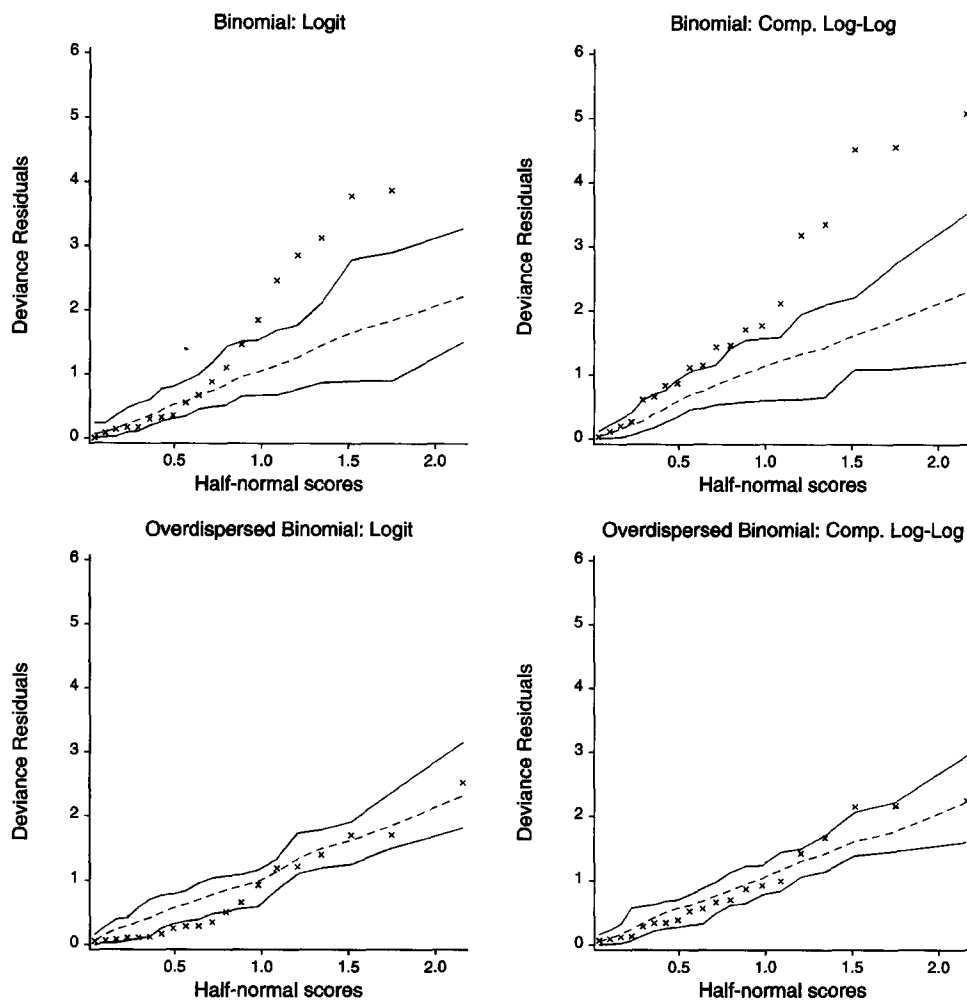
Fig. 3. Manly data. Half-normal plots: × – data; — simulated envelope.

parameter ϕ. For example, the constant overdispersion and beta-binomial variance function models differ only in the dependence of ϕ on the binomial sample size. Liang and McCullagh (1993) use plots of binomial residuals against sample size to suggest an appropriate model, however, it seems that such plots are rarely definitive. Ganio and Schafer (1992) also consider diagnostics for overdispersion models using a form of added variable plot to compare variance functions.

A useful omnibus technique for examining the residuals is to use a half-normal plot with a simulation envelope (Atkinson, 1985) which takes account of the overdispersion in the model. Demétrio and Hinde (1997) give a simple GLIM4 macro for such half-normal plots which is easily extended to a wide range of overdispersed models. Fig. 2 shows half-normal plots for some of the models considered in Table 1 for the orobanche data. In all cases the residuals are for the full `Species*Extract` model with a logit link function. This clearly shows the failure of the basic binomial

model and also suggests that the constant overdispersion (quasi-likelihood) model is inadequate. There is no clear evidence to choose between the other two variance functions, although the beta-binomial form seems most appropriate.

These plots can also indicate the failure of various other model aspects. For example, with the trout egg data, discussed in Section 2.2.2, Fig. 3 very clearly shows the failure of both ordinary binomial models and, more interestingly, the inadequacy of the overdispersed model with a logit link, while the complementary log-log link model has residuals inside of the envelope. The technique is also useful for detecting the presence of outliers which can have a large impact on overdispersion estimates.

Acknowledgements

References

Aitkin, M.A., 1995. NPML estimation of the mixing distribution in general statistical models with unobserved random variation. In: Seeber, G.U.H., Francis, B.J., Hatzinger, R., Steckel-Berger G. (Eds.), Statistical Modelling. Springer, New York.

Aitkin, M.A., 1996. A general maximum likelihood analysis of overdispersion in generalized linear models. Statist. Comput. 6, 251–262.

Aitkin, M.A., Francis, B.J., 1995. Fitting overdispersed generalized linear models by nonparametric maximum likelihood. GLIM Newslett. 25, 37–45.

Anderson, D.A., Hinde, J.P., 1988. Random effects in generalized linear models and the EM algorithm. Comm. Statist. A Theory Methods 17, 3847–3856.

Atkinson, A.C., 1985. Plots, Transformations and Regression. Clarendon Press, Oxford.

Breslow, N., 1984. Extra-Poisson variation in log-linear models. Appl. Statist. 33, 38–44.

Breslow, N., Clayton, D., 1993. Approximate inference in generalized linear mixed models. J. Amer. Statist. Assoc. 88, 9–25.

Brooks, R.J., 1984. Approximate likelihood ratio tests in the analysis of beta-binomial data, Appl. Statist. 33, 285–289.

Carroll, R.J., Ruppert, D., 1988. Transformation and Weighting in Regression, Chapman & Hall, London.

Clayton, D., 1996. Generalized linear mixed models. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J., (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Collett, D., 1991. Modelling Binary Data. Chapman & Hall, London.

Crowder, M., 1978. Beta-binomial anova for proportions. Appl. Statist. 27, 34–37.

Demétrio, C.G.B., Hinde, J.P., 1997. Half-normal plots and overdispersion. GLIM Newslett. 27, 19–26.

Efron, B., 1986. Double exponential families and their use in generalized linear regression. J. Amer. Statist. Assoc. 81, 709–721.

Engel, B., Keen, A., 1994. A simple approach for the analysis of generalized linear mixed models. Statist. Neerlandica 48, 1–22.

Finney, D.J., 1971. Probit Analysis, 3rd ed. Cambridge University Press, Cambridge.

Fitzmaurice, G.M., 1987. Model selection with overdispersed data. The Statistician 46, 81–91.

Ganio, L.M., Schafer, D.W., 1992. Diagnostics for overdispersion. J. Amer. Statist. Assoc. 87, 795–804.

Gaver, D.P., O'Muircheartaigh, I.G., 1987. Robust empirical Bayes analysis of event rates, Technometrics 29, 1–15.

Hinde, J.P., 1982. Compound Poisson regression models. In: Gilchrist, R. (Ed.), GLIM82. Springer, New York, 1982, pp. 109–121.

Hinde, J.P., 1996. Macros for fitting overdispersion models. GLIM Newslett. 26, 10–19.

Jorgensen, B., 1987. Exponential dispersion models (with discussion). J. Roy. Statist. Soc. Ser. B 49, 127–162.

Jorgensen, B., 1997. The Theory of Dispersion Models. Chapman & Hall, London.

Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. Canadian J. Statist. 15, 209–225.

Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models (with discussion). J. Roy. Statist. Soc. Ser. B 58, 619–678.

Liang, K.-Y., McCullagh, P., 1993. Case studies in binary dispersion. Biometrics 49, 623–630.

Lindsey, J.K., 1995. Modelling Frequency and Count Data. Clarendon Press, Oxford.

Manly, B., 1978. Regression models for proportions with extraneous variance. Biometrie-Praximetrie 18, 1–18.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall, London.

McCulloch, C., 1994. Maximum likelihood variance components estimation for binary data. J. Amer. Statist. Assoc. 89, 330–335.

Moore, D.F., 1986. Asymptotic properties of moment estimates for overdispersed counts and proportions. Biometrika 73, 583–588.

Nelder, J.A., 1985. Quasi-likelihood and GLIM. In: Gilchrist, R., Francis, B., Whittaker, J. (Eds.), Generalized Linear Models. Springer, Berlin, pp. 120–127.

Nelder, J.A., Y. Lee, Y., 1992. Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons. J. Roy. Statist. Soc. Ser. B 54, 273–284.

Nelder, J.A., Pregibon, D., 1987. An extended quasi-likelihood function. Biometrika 74, 221–232.

Prentice, R.L., 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. J. Amer. Statist. Assoc. 81, 321–327.

Weil, C., 1970. Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. Food Cosmetics Toxicol. 8, 177–182.

Williams, D.A., 1982. Extra-binomial variation in logistic-linear models. Appl. Statist. 31, 144–148.

Williams, D.A., 1996. Overdispersion in logistic-linear models. In: Morgan, B.J.T. (Ed.), Statistics in Toxicology. Clarendon Press, Oxford.