Overdispersion: Models and Estimation

A Short Course for SINAPE 1998

John Hinde
MSOR Department, Laver Building, University of Exeter,
North Park Road, Exeter, EX4 4QE, UK
Email: J.P.Hinde@exeter.ac.uk
Fax: +44 1392 264460

Clarice G.B. Demétrio
Departamento de Matemática e Estatística, ESALQ, USP
Caixa Postal 9
13418-900 Piracicaba, SP
Email: Clarice@carpa.ciagri.usp.br
Fax: 019 4294346

April 12, 2007

Preface

These notes are for the short course to be given at the 13th Brazilian Symposium of Probability and Statistics (13^o SINAPE), Caxambu, Minas Gerais, Brazil in July 1998.

The notes are intended to give a simple introduction to the idea of overdispersion, present some commonly used models and discuss general ideas of estimation. The literature on overdispersion is very large and, in the time available, we can only present a limited overview of the subject, reflecting our own particular interests. However, we have attempted to include links to those aspects of the literature that are not covered in great detail and we hope that these notes will form both a useful summary of the basic ideas and starting point for anyone involved with overdispersed data.

Acknowledgements

The authors are grateful to FAPESP and The Royal Society for funding a number of exchange visits.

Contents

Chapter 1

Introduction

The basic context for the material in this short course is that of *generalized linear models*. These were first introduced by Nelder and Wedderburn (1972) as an extension to the standard normal theory linear model. The assumption of normality is relaxed to the exponential family of distributions and the resulting unified theory includes the standard techniques of log-linear modelling for count data and the logistic model for proportions. A full discussion of generalized linear models is given in McCullagh and Nelder (1989) while Dobson (1990) provides a simple introduction. Aitkin et al. (1989) give a thorough account of the practical application of generalized linear models using the software package GLIM, which was specifically written for the fitting of generalized linear models (Francis et al. (1993) for the details on the current version). Other accounts on the application and extension of generalized linear models include Firth (1991), Lindsey (1989, 1995, 1997) and Fahrmeir and Tutz (1994).

Generalized linear models are applicable when we have a single response variable Y and associated explanatory variables $x_1, x_2, \ldots, x_p$, where typically $x_1 \equiv 1$ to include the usual constant term in the model. For a random sample of n observations $(y_i, \mathbf{x}_i)$, where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T$ is the column vector of explanatory variables, the three components of a generalized linear model are:

- independent random variables Y_i, $i = 1, \ldots, n$, from a linear exponential family distribution with means μ_i and constant scale parameter ϕ, i.e.

observations from a density of the form

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\} \tag{1.1}$$

where $\mu = \mathbf{E}[Y] = b'(\theta)$ and $\mathrm{Var}(Y) = \phi b''(\theta)$.
θ is commonly called the canonical, or natural, parameter;

- a linear predictor vector $\boldsymbol{\eta}$ given by

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

where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ is the $n \times p$ design matrix;

- a link function $g(\cdot)$ relating the mean to the linear predictor, i.e.

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$$

If we choose the link function so that $g(\mu) = \theta$, the linear predictor is directly modelling the canonical parameter and such a link function is referred to as the canonical, or natural, link. This frequently results in a sensible scale for the modelling with useful practical interpretations of the regression parameters; there are also theoretical advantages, in terms of the existence of a set of simple sufficient statistics for the regression parameters $\boldsymbol{\beta}$ and some simplification of the computational algorithm.

Maximum likelihood estimates for the regression parameters are most easily obtained using Fisher-scoring which can be implemented as as IWLS (Iteratively Weighted Least-Squares) procedure, see McCullagh and Nelder (1989).

1.1 Models for proportion and count data

In this course we will restrict our attention to two simple modelling situations where overdispersion is particularly common. The first is the modelling of proportions, where the usual starting point is the binomial regression model. The second concerns the analysis of count data and the Poisson regression model. These are both specific examples of generalized linear models (glms)

and hence our focus on this class of models. Of course, the concept of overdispersion is more widely applicable than this, but the family of glms and these particular examples are sufficient for introducing and illustrating the key ideas and methods.

1.1.1 Binomial regression models

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

$$\mathbf{E}[Y_i] = \mu_i = m_i \pi_i,$$

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

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

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

$$\text{Var}(Y_i) = m_i \pi_i (1 - \pi_i). \tag{1.2}$$

The canonical link function for the binomial distribution is the logit link

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

which corresponds to modelling on the log-odds scale. 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.

1.1.2 Poisson regression models

If we now suppose 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

$$\text{Var}(Y_i) = \mu_i. \tag{1.3}$$

Here the canonical link function is the log link

$$g(\mu_i) = \log(\mu_i) = \eta_i,$$

providing a multiplicative model for the Poisson rate parameter μ_i.

In a number of practical situations where we wish to model count data we may have different observation periods for our counts. If we are interested in modelling the underlying Poisson rates, λ_i, we have to take account of these different observation times, or exposure periods, t_i. The Poisson model for the observed counts is now $Y_i \sim \text{Pois}(t_i\lambda_i)$ and taking a log-linear model for the rates, $\log(\lambda_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ results in the following log-linear model for the Poisson means

$$\log(\mu_i) = \log(t_i\lambda_i) = \log(t_i) + \mathbf{x}_i^T\boldsymbol{\beta},$$

where the $\log(t_i)$ is included as a fixed term, or *offset*, in the model. Models of this form are frequently used in occupational disease exposure studies, dilution assays, machine reliability modelling and insurance work.

1.2 Overdispersion: causes and consequences

1.2.1 Overdispersion in glms

When applying generalized linear models with a known scale parameter, as is certainly the case for the binomial and Poisson distributions where $\phi = 1$, subject to certain asymptotic conditions for a well fitting model we would expect

$$\text{Residual Deviance} \approx \text{Residual degrees of freedom (df)}$$

What if the Residual Deviance $\gg$ Residual df? There are two possible scenarios we need to consider.

(i) We may simple have a badly fitting model for one of a number of reasons such as

- omitted terms or variables in the linear predictor;
- incorrect relationship between mean and explanatory variables, *i.e.* we may have the wrong link function or need to transform one or more explanatory variables;
- outliers.

Standard model diagnostics allow us to explore these aspects, see McCullagh and Nelder (1989), Chapter 12.

(ii) The variation may simply be greater than that predicted by model and it is this phenomenon that is described as ***overdispersion***. In essence our model is too restrictive for our data and for the models discussed in Section 1.1, rather than the variance identities given by equations 1.2, 1.3, we have

- proportion data with $\text{Var}(Y_i) > m_i\pi_i(1-\pi_i)$;
- count data with $\text{Var}(Y_i) > \mu_i$.

1.2.2 Causes of overdispersion

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

- variability of experimental material – this can be thought of as individual variability of the experimental units and may give an additional component of variability which is not accounted for by the basic model;
- correlation between individual responses – for example in cancer studies involving litters of rats we may expect to see some correlation between rats in the same litter;
- cluster sampling;
- aggregate level data – the aggregation process can lead to compound distributions;

- omitted unobserved variables – in some sense the other categories are all special cases of this, but generally in a rather complex way.

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 so in general it is difficult to infer the precise cause, or underlying process, leading to the overdispersion.

1.2.3 Consequences of overdispersion

When we identify the possible presence of overdispersion, what are the consequences of failing to take it into account? Firstly, the standard errors obtained from the model will be incorrect and may be seriously underestimated and consequently we may incorrectly assess the significance of individual regression parameters. Also, changes in deviance associated with model terms will also be too large and this will lead to the selection of overly complex models. Finally, our interpretation of the model will be incorrect and any predictions will be too precise.

1.3 Examples

1.3.1 *Example: Germination of Orobanche seed*

Table 1.1. presents data from a study of the germination of two species of *Orobanche* seeds. The seeds were grown on 1/125 dilutions of two different root extract media (cucumber or bean) in a 2×2 factorial layout with replicates. The data (y_i/m_i) consist of the number of seeds, m_i, and the number germinating, y_i, for each replicate. Interest focuses on the possible differences in germination rates for the two types of seed and root extract and whether there is any interaction. The data were first analysed by Crowder (1978) and are also discussed in Collett (1991).

Fitting the standard binomial logit model to these data gives a deviance change for the `Species.Extract` interaction term of 6.41 on 1 degree of freedom (df). Comparing this with a χ^2_1 distribution suggests that the interaction term is significant, as $\chi^2_{1;\,0.95} = 3.84$. However, the residual deviance for

Table 1.1: Orobanche seed germination data (table entries y_i/m_i).

O. aegyptiaca 75		*O. aegyptiaca 73*	
Bean	Cucumber	Bean	Cucumber
10/39	5/6	8/16	3/12
23/62	53/74	10/30	22/41
23/81	55/72	8/28	15/30
26/51	32/51	23/45	32/51
17/39	46/79	0/4	3/7
	10/13		

this model is 33.28 on 17 df giving some evidence of possible overdispersion ($\chi^2_{17;\,0.99} = 33.41$). If we were to take account of this overdispersion, would this change our inference about the significance of the interaction and would the resulting model provide a better description of the data?

A possible explanation for overdispersion is that the experimental conditions of the replicates (individual cultivation plates) may not be identical. Thus we can think of some additional variation of the probability of germination over the replicates and we could consider taking account of this in the modelling process by allowing the probabilities to vary over replicates. There are of course a number of different possible causes for this variation; it may simply be intrinsic variability in the experimental material, the growth media, or, if it were due to slightly different experimental conditions, we could in principle think of taking additional measurements on say temperature or light conditions of each replicate and including these as additional explanatory variables. In this case we could consider our overdispersion as arising due to omitted unobserved variables. Another way of thinking of the effect of this additional variation over the replicates is that the responses for the individual seeds within a replicate will exhibit some degree of correlation due to the unknown common factors. Alternatively, a possible mechanism for correlation between individual seeds is that the germination of one seed may release a

chemical substance which induces germination of the other seeds. We will consider these different approaches to overdispersion for proportion data in Chapter 2.

1.3.2 *Example: Worldwide airline fatalities*

Table 1.2: Worldwide airline fatalities, 1976-85.

Year	Fatal accidents	Passenger deaths	Passenger miles (100 million)
1976	24	734	3863
1977	25	516	4300
1978	31	754	5027
1979	31	877	5481
1980	22	814	5814
1981	21	362	6033
1982	26	764	5877
1983	20	809	6223
1984	16	223	7433
1985	22	1066	7107

The data in Table 1.2 relate to airline fatalities and to determine whether air travel is becoming safer we can use Poisson regression models to fit a time trend to the accident statistics. The recorded passenger miles, m_i, indicate that in some sense the volume of air traffic nearly doubled over the 10 year period 1976-1985. We clearly need to take account of this fact in any models that we fit to these data. We can use this variable as an exposure variable in a Poisson log-linear model as in Section 1.1.2; it will act as a surrogate for the volume of air traffic. We can summarize the models of interest as:

$$\begin{aligned} Y_i \sim &\quad \text{Pois}(m_i\lambda_i) \\ \log \lambda_i = &\quad \beta_0 + \beta_1 \texttt{year}_i \end{aligned}$$

that is

- Passenger miles (m_i) as exposure variable
- Poisson log-linear model
- Linear time trend

Taking Y as the numbers of fatal accidents gives us analysis of flight accident rate over time, while taking Y as the number of passenger deaths focuses on our own individual concern, the passenger death rate. The results for these two analyses are summarized below:

Fatal Accidents:
Deviance(`time trend`) = 20.68 on 1 df
Residual Deviance = 5.46 on 8 df

Passenger deaths:
Deviance(`time trend`) = 202.1 on 1 df
Residual Deviance = 1051.5 on 8 df

In both cases it seems that the time trend is highly significant and, in fact, the estimate β_1 is negative in both cases indicating that air travel was becoming safer over the period in question. However, while the residual deviance for the fatal accident model indicates a good fit, the residual deviance for the passenger death model is very large compared to the degrees of freedom. This large degree of overdispersion is due to compounding with the aircraft size, as each fatal accident leads to some multiple number of deaths. So, while the fatal accident data is consistent with an underlying Poisson process giving Poisson counts for the number of fatal accidents each year, the number of passenger deaths each year is a Poisson sum of random variables from the aircraft size distribution, which is not Poisson distributed. We will discuss this further in Chapter 3.

Chapter 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. For binomial data, Collett (1991) gives a good practical introduction to some of these 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).

2.1 Mean-variance models

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.

2.1.1 Proportion data

As in Section 1.1.1 we suppose that the random variables Y_i represent counts of successes out of samples of size m_i, $i = 1, \ldots, n$.

Constant overdispersion

A constant overdispersion model replaces 1.2 by

$$\mathrm{Var}(Y_i) = \phi m_i \pi_i (1 - \pi_i). \tag{2.1}$$

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).

A general variance function

A more general variance function will allow the overdispersion to depend upon both m_i and π_i. A set of GLIM4 macros ((Hinde 1996)) allows the fitting of overdispersed binomial data with a variance function $\text{Var}(Y_i) = m_i\pi_i(1-\pi_i)[1+\phi f(m_i,\pi_i)]$, where f is a function specified by the user. One particular general form, which will be seen to include many of the standard variance functions for proportion data is

$$\text{Var}(Y_i) = m_i\pi_i(1-\pi_i)\left[1+\phi(m_i-1)^{\delta_1}\{\pi_i(1-\pi_i)\}^{\delta_2}\right]. \tag{2.2}$$

The standard binomial model corresponds to $\phi = 0$, while $\delta_1 = \delta_2 = 0$ gives the constant overdispersion model in a slightly different parameterization; other values of δ_1 and δ_2, lead to variance functions which we will meet in Section 2.2.1. (Specifically $\delta_1 = 1$, $\delta_2 = 0$ gives the beta-binomial variance function and $\delta_1 = \delta_2 = 1$, the Williams type III model.)

2.1.2 Count data

As in Section 1.1.2 we suppose that the random variables Y_i represent counts with means μ_i.

Constant overdispersion

A constant overdispersion model replaces 1.3 by

$$\text{Var}(Y_i) = \phi\mu_i. \tag{2.3}$$

A variance function of this form can arise through a simple compounding process of taking a Poisson random sum of independent and identically distributed (iid) random variables as in Example 1.3.2, the airline fatality data. To see this suppose that $N \sim \text{Pois}(\mu_N)$ and $T = \sum_{i=1}^{N} X_i$, where X_i are iid random variables. Using the standard results for conditional expectation and variance

we obtain

$$\begin{aligned}\mathbf{E}[T] &= \mu_T = \mathbf{E}_N(\mathbf{E}[T|N]) = \mathbf{E}_N[N\mu_X] = \mu_N\mu_X \\ \\ \mathrm{Var}(T) &= \mathbf{E}_N[\mathrm{Var}(T|N)] + \mathrm{Var}_N(\mathbf{E}[T|N]) \\ &= \mu_T\left(\frac{\sigma_X^2}{\mu_X} + \mu_X\right) = \mu_T\frac{\mathbf{E}[X^2]}{\mathbf{E}[X]}\end{aligned}$$

that is $\mathrm{Var}(T) = \phi\mu_T$, where ϕ depends upon the first two moments of the X-distribution and we have overdispersion if $\mathbf{E}[X^2] > \mathbf{E}[X]$. Note however that there are other overdispersion processes that can lead to a constant overdispersion model. (In this type of model it is also possible to have underdispersion, although this is not very common in practice. If the X_i are Bernoulli random variables, $\mathbf{E}[X^2] = \mathbf{E}[X]$, and we have no overdispersion; this corresponds to a thinned Poisson process and T is also a Poisson random variable.)

A general variance function

A general variance function which encompasses the various commonly used models is

$$\mathrm{Var}(Y_i) = \mu_i\left\{1 + \phi\mu_i^{\delta}\right\}, \tag{2.4}$$

although other natural extensions would be to consider a general quadratic variance function or a simple power function. The standard Poisson model corresponds to taking $\phi = 0$ in (2.4), while $\delta = 0$ gives the constant overdispersion model in a slightly different parameterization; taking $\delta = 1$ leads to the negative binomial variance function discussed in Section 2.2.2.

2.2 Two-stage models

2.2.1 Binomial data

Beta-binomial Variance Function (Williams Type II)

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

then, unconditionally, we have

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

and

$$\mathrm{Var}(Y_i) = m_i\pi_i(1-\pi_i)[1+\phi(m_i-1)]. \tag{2.5}$$

A special case of this is the beta-binomial distribution, which is obtained by assuming that the P_i's have Beta(α_i, β_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. Estimation will be discussed in more detail in Chapter 3.

A conceptually different model for overdispersion is to assume that the individual binary responses are not independent. Writing $Y_i = \sum_{j=1}^{m_i} R_{ij}$, where R_{ij} are Bernoulli random variables with $\mathbf{E}[R_{ij}] = \pi_i$ and $\mathrm{Var}(R_{ij}) = \pi_i(1-\pi_i)$, then, assuming a constant correlation ρ between the R_{ij}'s for $j \neq k$, we have $\mathrm{Cov}(R_{ij}, R_{ik}) = \rho\pi_i(1-\pi_i)$ and

$$\begin{aligned}
\mathbf{E}[Y_i] &= m_i\pi_i \\
\mathrm{Var}(Y_i) &= \sum_{j=1}^{m_i}\mathrm{Var}(R_{ij}) + \sum_{j=1}^{m_i}\sum_{k\neq j}\mathrm{Cov}(R_{ij},R_{ik}) \\
&= m_i\pi_i(1-\pi_i) + m_i(m_i-1)[\rho\pi_i(1-\pi_i)] \\
&= m_i\pi_i(1-\pi_i)[1+\rho(m_i-1)],
\end{aligned}$$

which is of exactly the same form as (2.5). 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)

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

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

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). 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.

Considered as the two-stage model, the $\text{logit}(P_i)$ have a normal distribution with variance σ^2, i.e. $\text{logit}(P_i) \sim \text{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$. Writing

$$U_i = \text{logit}(P_i) = \log\frac{P_i}{(1-P_i)} \Rightarrow P_i = \frac{e^{U_i}}{(1+e^{U_i})}$$

and using Taylor series for P_i, around $U_i = \mathbf{E}[U_i] = \mathbf{x}_i^T\boldsymbol{\beta}$, we have

$$P_i = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})} + \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})^2}(U_i - \mathbf{x}_i^T\boldsymbol{\beta}) + o(U_i - \mathbf{x}_i^T\boldsymbol{\beta}).$$

Then

$$\mathbf{E}(P_i) \approx \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})} := \pi_i$$

and

$$\text{Var}(P_i) \approx \left[\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}})^2}\right]^2 \text{Var}(U_i) = \sigma^2\pi_i^2(1-\pi_i)^2$$

Consequently the variance function for the logistic-normal model can be approximated by

$$\text{Var}(Y_i) \approx m_i\pi_i(1-\pi_i)[1+\sigma^2(m_i-1)\pi_i(1-\pi_i)], \tag{2.6}$$

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

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.

2.2.2 Count data

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 $\mathbf{E}(\theta_i) = \mu_i$ and $\text{Var}(\theta_i) = \sigma_i^2$. Unconditionally, we have $\mathbf{E}(Y_i) = \mu_i$ 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 $\mathbf{E}(Y_i) = k/\lambda_i = \mu_i$ and

$$\text{Var}(Y_i) = \mu_i + \mu_i^2/k. \tag{2.7}$$

For fixed values of k this distribution is in the exponential family and so we are still in the generalized linear modelling framework.

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 $\text{Var}(Y_i) = \mu_i + \mu_i/\lambda \equiv \phi\mu_i$, the constant overdispersion model. However, the resulting distribution for Y_i is not in the exponential family, see Nelder and Lee (1992) for details of maximum likelihood estimation

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. 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

$$\begin{aligned}
\mathbf{E}[Y_i] &= \mathbf{E}_{Z_i}\left(\mathbf{E}[Y_i|Z_i]\right) = \mathbf{E}_{Z_i}[e^{\mathbf{x}_i^T\boldsymbol{\beta}+\sigma Z_i}] = e^{\mathbf{x}_i^T\boldsymbol{\beta}+\frac{1}{2}\sigma^2} := \mu_i \\
\text{Var}(Y_i) &= \mathbf{E}_{Z_i}[\text{Var}(Y_i|Z_i)] + \text{Var}_{Z_i}(\mathbf{E}[Y_i|Z_i]) \\
&= e^{\mathbf{x}_i^T\boldsymbol{\beta}+\frac{1}{2}\sigma^2} + \text{Var}_{Z_i}(e^{\mathbf{x}_i^T\boldsymbol{\beta}+\sigma Z_i}) \\
&= e^{\mathbf{x}_i^T\boldsymbol{\beta}+\frac{1}{2}\sigma^2} + e^{2\mathbf{x}_i^T\boldsymbol{\beta}+\sigma^2}(e^{\sigma^2}-1).
\end{aligned}$$

So the variance function for this model is of the form $\text{Var}(Y_i) = \mu_i + k'\mu_i^2$, that is, the same as for the negative binomial distribution. In fact, with a log-link function and an additive continuous 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 the same as those for the negative binomial distribution. Alternatively, by using Aitkin's NPML method we can avoid any specific distributional assumption for the random effect.

Chapter 3

Estimation methods

In this Chapter we will consider various approaches which can be used for the estimation of overdispersion models. The use of a particular method is bound up with the specification of the model. For example, full maximum likelihood is only available if we have a complete specification of the probability model. If we have only specified some more general form of the variance function, then we need to use an estimation method that only depends upon the first two moments, such as quasi-likelihood and related methods.

In the literature a wide variety of different estimation methods have been proposed. Here the aim is to set out many of these different methods and see the relationships between them.

3.1 Maximum likelihood

Full maximum likelihood estimation can be used whenever we have a model which gives a complete and explicit probability model for the response variable. For proportion data this is the case with the beta-binomial distribution and for count data when we have a negative binomial distribution. For models with a random effect in the linear predictor, although we have a fully specified probability model, in general it has no simple closed form and we need to use the EM-algorithm, see Dempster, Laird, and Rubin (1977).

3.1.1 Beta-binomial distribution

If we assume that $Y_i|P_i \sim \text{Bin}(m_i, P_i)$, where $P_i \sim \text{Beta}(\alpha_i, \beta_i)$, $i = 1, \ldots, n$, that is

$$f_{Y_i|P_i}(y_i|p_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i}, \qquad y_i = 0, 1, \ldots, m_i$$

and

$$f_{P_i}(p_i) = \frac{p_i^{\alpha_i-1}(1-p_i)^{\beta_i-1}}{B(\alpha_i, \beta_i)}, \qquad 0 \le p_i \le 1$$

then, unconditionally

$$\begin{aligned} f_{Y_i}(y_i) &= \binom{m_i}{y_i} \frac{B(\alpha_i + y_i, m_i + \beta_i - y_i)}{B(\alpha_i, \beta_i)} \\ &= \frac{m_i!}{y_i!(m_i - y_i)!} \frac{\Gamma(\alpha_i + y_i)\ \Gamma(m_i + \beta_i - y_i)}{\Gamma(m_i + \alpha_i + \beta_i)} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\ \Gamma(\beta_i)} \end{aligned}$$

and

$$\mathbf{E}(Y_i) = \mathbf{E}_{P_i}[\mathbf{E}_{Y_i|P_i}(Y_i)] = \mathbf{E}_{P_i}[m_i P_i] = m_i \frac{\alpha_i}{\alpha_i + \beta_i} := m_i \pi_i$$

$$\begin{aligned} \text{Var}(Y_i) &= \mathbf{E}_{P_i}[\text{Var}_{Y_i|P_i}(Y_i)] + \text{Var}_{P_i}[\mathbf{E}_{Y_i|P_i}(Y_i)] \\ &= \mathbf{E}_{P_i}[m_i P_i(1-P_i)] + \text{Var}_{P_i}(m_i P_i) \\ &= m_i \pi_i (1-\pi_i) \left[1 + (m_i - 1)\frac{1}{\alpha_i + \beta_i + 1}\right]. \end{aligned}$$

In applying the beta-binomial model it is common practice to treat $\alpha_i + \beta_i = c$, i.e. constant over i. This corresponds to taking

$$\text{Var}(P_i) = \frac{\pi_i(1-\pi_i)}{\alpha_i + \beta_i + 1} \propto \pi_i(1-\pi_i)$$

and using a fixed effective sample size for the P_i's. Writing

$$\phi = 1/(c+1) = 1/(\alpha_i + \beta_i + 1)$$

gives

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

as in equation (2.5). (It is clear from these derivations that taking P_i to be any random variable with the same mean and variance structure will give the same mean-variance relationship for Y_i.) Taking $c = \infty$ ($\phi = 0$) corresponds to the beta random variables being constants and gives the standard binomial distribution.

Ignoring constants the log-likelihood can be written as

$$\begin{aligned}\ell(\boldsymbol{\pi}, c|\mathbf{y}) &= \sum_{i=1}^{n}\left\{\sum_{r=0}^{y_i-1}\log(c\pi_i + r) + \sum_{s=0}^{m_i-y_i-1}\log(c(1-\pi_i)+s)\right. \\ &\qquad \left. - \sum_{t=0}^{m_i-1}\log(c+t)\right\}\end{aligned}$$

or, more simply, as

$$\begin{aligned}\ell(\boldsymbol{\pi}, c|\mathbf{y}) &= \sum_{i=1}^{n}\Big\{\log\Gamma(c\pi_i + y_i) - \log\Gamma(c\pi_i) + \log\Gamma[c(1-\pi_i) + m_i - y_i] \\ &\qquad - \log\Gamma[c(1-\pi_i)] - \log\Gamma(m_i + c) + \log\Gamma(c)\Big\} \\ &= \sum_{i=1}^{n}\big\{\mathrm{dlg}(y_i, c\pi_i) + \mathrm{dlg}[m_i - y_i, c(1-\pi_i)] - \mathrm{dlg}(m_i, c)\big\}\end{aligned}$$

where

$$\mathrm{dlg}(y, a) = \log\Gamma(y + a) - \log\Gamma(a)$$

a difference of $\log\Gamma$ functions.

Modelling the π_i's with a linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ and link function $g(\pi_i) = \eta_i$ we obtain the following score equations for maximum likelihood estimation:

$$\begin{aligned}\frac{\partial\ell}{\partial\beta_j} &= c\sum_{i=1}^{n}\Big\{\mathrm{ddg}(y_i, c\pi_i) - \mathrm{ddg}\big(m_i - y_i, c[1-\pi_i]\big)\Big\}\frac{1}{g'(\pi_i)}x_{ij} \\ \frac{\partial\ell}{\partial c} &= \sum_{i=1}^{n}\Big\{\pi_i\,\mathrm{ddg}(y_i, c\pi_i) + (1-\pi_i)\,\mathrm{ddg}\big(m_i - y_i, c[1-\pi_i]\big) - \mathrm{ddg}(m_i, c)\Big\}\end{aligned}$$

where

$$\mathrm{ddg}(y,a) = \frac{\partial}{\partial a}\Big(\mathrm{dlg}(y,a)\Big) = \psi(y+a) - \psi(a)$$

and ψ is the di-gamma function.

One obvious strategy for solving these equations is to use a Gauss-Seidel (successive relaxation) algorithm iterating over the following steps:

(i) for fixed c, solve $\partial\ell/\partial\boldsymbol{\beta} = 0$ to obtain an updated estimate for $\boldsymbol{\beta}$ and hence $\boldsymbol{\pi}$;

(ii) for fixed $\boldsymbol{\pi}$ (*i.e.* $\boldsymbol{\beta}$), solve $\partial\ell/\partial c = 0$ to obtain an updated estimate for c.

Step (ii) is easily implemented using a simple Newton-Raphson type procedure; the second partial derivatives just involve tri-gamma functions. Step (i) can also be implemented using Newton-Raphson or Fisher scoring, however, because even for fixed c the beta-binomial model is not in the linear exponential family, the Fisher scoring procedure does not reduce to a simple generalized linear model fitting algorithm.

One simple modification is to replace the maximum likelihood estimating equations for $\boldsymbol{\beta}$ by quasi-likelihood equations, see Section 3.2. Initial estimates can be obtained by

- fitting a standard binomial model to give initial values for $\boldsymbol{\pi}$;
- setting $c_0 = (1-\phi_0)/\phi_0$ where

$$\phi_0 = \frac{X^2 - (n-p)}{\sum_i (m_i - 1)\left[1 - m_i\widehat{\pi}_i(1-\widehat{\pi}_i)h_i\right]}$$

  and $h_i = \mathrm{Var}(\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i)$ is the variance of the linear predictor and X^2 is the Pearson X^2 statistic from the binomial fit. This comes from equating the X^2 statistic from the binomial fit to its expected value under the beta-binomial model.

3.1.2 Negative binomial distribution

If the random variables Y_i represent counts with means θ_i and we specify a a two-stage model with $Y_i \sim \text{Pois}(\theta_i)$, where the θ_i's are random variables with a $\Gamma(k, \lambda_i)$ distribution, this leads to a negative binomial distribution for the Y_i with

$$f_{Y_i}(y_i; \mu_i, k) = \frac{\Gamma(k + y_i)}{\Gamma(k) y_i!} \frac{\mu_i^{y_i} k^k}{(\mu_i + k)^{k+y_i}}, \qquad y_i = 0, 1, \ldots$$

and $\mathbf{E}(Y_i) = k/\lambda_i = \mu_i$

$$\begin{aligned} \text{Var}(Y_i) &= \mathbf{E}_{\theta_i}[\text{Var}(Y_i|\theta_i)] + \text{Var}_{\theta_i}(\mathbf{E}[Y_i|\theta_i]) \\ &= \mathbf{E}[\theta_i] + \text{Var}(\theta_i) = \frac{k}{\lambda_i} + \frac{k}{\lambda_i^2} \\ &= \mu_i + \mu_i^2/k. \end{aligned}$$

Under the negative-binomial model we have the following expression for the log-likelihood:

$$\begin{aligned} \ell(\boldsymbol{\mu}, k; \mathbf{y}) &= \sum_{i=1}^{n} \Big\{ y_i \log \mu_i + k \log k - (k + y_i) \log(k + \mu_i) \\ &\qquad + \log \frac{\Gamma(k + y_i)}{\Gamma(k)} - \log y_i! \Big\} \\ &= \sum_{i=1}^{n} \Big\{ y_i \log \mu_i + k \log k - (k + y_i) \log(k + \mu_i) \\ &\qquad + \text{dlg}(y_i, k) - \log y_i! \Big\} \end{aligned}$$

Notice here that for fixed values of k we have a linear exponential family model and consequently a generalized linear model.

Modelling the μ_i's with a linear predictor $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ and link function $g(\mu_i) = \eta_i$ we obtain the following score equations for maximum likelihood

estimation:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n}\left\{\frac{y_i}{\mu_i}-\frac{k+y_i}{k+\mu_i}\right\}\frac{\partial \mu_i}{\partial \beta_j}\\ &= \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\mu_i(1+\frac{\mu_i}{k})}\frac{1}{g\prime(\mu_i)}x_{ij}\\ \frac{\partial \ell}{\partial k} &= \sum_{i=1}^{n}\left\{\mathrm{ddg}(y_i,k)-\log(\mu_i+k)-\frac{k+y_i}{k+\mu_i}+\log k+1\right\}.\end{aligned}$$

The score equations for $\boldsymbol{\beta}$ are the usual quasi-score equations for a glm with $V(\mu)=\mu(1+\frac{\mu}{k})$ and $g(\boldsymbol{\mu})=\eta$ and so provide a simple approach for fitting negative binomial regression models using a Gauss-Seidel approach and iterating over the following steps:

(i) For a fixed value of k, estimate $\boldsymbol{\beta}$ using a standard glm fit (IRLS) with a variance function $V(\mu)=\mu+\mu^2/k$;

(ii) for fixed $\boldsymbol{\beta}$, and hence $\boldsymbol{\mu}$, estimate k using a Newton-Raphson iterative scheme

$$k^{(m+1)}=k^{(m)}-\left(\frac{\partial \ell}{\partial k}\Big/\frac{\partial^2 \ell}{\partial k^2}\right)\Bigg|_{k^{(m)}}$$

and iterating between them until convergence. The second derivative with respect to k is given by

$$\frac{\partial^2 \ell}{\partial k^2}=\sum_{i=1}^{n}\left\{\mathrm{dtg}(y_i,k)-\frac{1}{\mu_i+k}+\frac{k+y_i}{(k+\mu_i)^2}-\frac{1}{\mu_i+k}+\frac{1}{k}\right\}$$

where $\mathrm{dtg}(y,k)=\partial\{\mathrm{ddg}(y,k)\}/\partial k$ is a difference of tri-gamma functions. Also

$$\frac{\partial^2 \ell}{\partial \beta_j \partial k} = \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{(k+\mu_i)^2}\frac{1}{g\prime(\mu_i)}x_{ij}$$

and so $\mathbf{E}\left(\dfrac{\partial^2 \ell}{\partial \beta_j \partial k}\right)=0$, that is k and β_j are asymptotically uncorrelated. This asymptotic independence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{k}$ means that the standard errors for $\boldsymbol{\beta}$ from the IRLS fit are correct, see Lawless (1987) for details.

Initial values for k can be obtained from

- fitting a standard Poisson model to obtain $\widehat{\mu}_i$
- set

$$k_0 = \frac{\sum_{i=1}^{n} \widehat{\mu}_i(1 - h_i\widehat{\mu}_i)}{\sum_{j=1}^{n} \dfrac{(y_j - \widehat{\mu}_j)^2}{\widehat{\mu}_j} - (n-p)}$$

  where $h_i = \mathrm{Var}(\widehat{\boldsymbol{\beta}}^T \mathbf{x}_i)$ is the variance of the linear predictor. This comes from equating the Pearson X^2 statistic from the Poisson fit to its expected value under the negative binomial model, see Breslow (1984).

3.1.3 Random effect in the linear predictor

If we assume that the linear predictor, η_i, has some continuous distribution and this distribution is taken to be in the location-scale family then this corresponds to including an additive random effect in the linear predictor. 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 for proportion data and the Poisson-normal model for count data. 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) for details of estimation. A general approach to the estimation of models of this type is to use the EM-algorithm (Dempster, Laird, and Rubin 1977) with Gaussian quadrature to integrate over the normal distribution, following the same approach given by Hinde (1982) for the Poisson distribution.

To illustrate this, we will consider the logistic-normal model. Here we have

$$Y_i \sim \mathrm{Bin}(m_i, \pi_i), \quad \mathrm{logit}(\pi_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma Z_i \quad \text{and} \quad Z_i \sim N(0,1).$$

To apply the EM-algorithm we need to maximize the expected of complete data log-likelihood, where the expectation is with respect to the conditional distribution of the unobserved variable $\mathbf{z}$, *i.e.* the distribution of $\mathbf{z}$ given $\mathbf{y}$

and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$. The E-step involves finding

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^c) &= \mathbf{E}[\log f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^c] \\ &= \mathbf{E}[\log f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^c] + \mathbf{E}[\log \boldsymbol{\phi}(\mathbf{z})|\mathbf{y}, \boldsymbol{\theta}^c] \end{aligned}$$

where $\boldsymbol{\theta}^c$ is some current parameter estimate and $\phi()$ is the standard normal pdf. The expectations are not analytically tractable but can be evaluated using K-point quadrature with quadrature points $z_1, z_2, \ldots, z_K$ and associated weights $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_K$. So we can approximate Q by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^c) \approx \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij} \log f(y_i|z_j, \boldsymbol{\theta}) + \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij} \log \phi(z_j)$$

where

$$w_{ij} = \frac{\varepsilon_i f(y_i, z_j|\boldsymbol{\theta}^c)}{\sum_{l=1}^{K} \varepsilon_l f(y_i, z_l|\boldsymbol{\theta}^c)}$$

are the "posterior" probabilities of z_j given y_i and $\boldsymbol{\theta}^c$.

In terms of estimating the unknown parameters $\boldsymbol{\theta}$, Q has a particularly simple form being a weighted sum of binomial log-likelihoods. Consequently, updated estimates for $\boldsymbol{\beta}$ and σ can be obtained by fitting a logit model to expanded data (K-copies of the original data) with

- a y-variate $\mathbf{y}^* = (\mathbf{y}^T, \mathbf{y}^T, \ldots, \mathbf{y}^T)^T$
- explanatory variables for y_{ij}^*: $\mathbf{x}_i$ and z_j
- weights w_{ij}

Iterating the E and M steps provides maximum likelihood estimates for $\boldsymbol{\beta}$ and σ.

3.2 Maximum quasi-likelihood

Constant overdispersion models as in (2.1) and (2.3) fit into the class of simple quasi-likelihood models as described by Wedderburn (1974). The principle

here for a model with variance of the form $\mathrm{Var}(Y_i) = \phi V_i(\mu_i)$ is to estimate the regression parameters to minimize the quasi-likelihood

$$Q = -\tfrac{1}{2}\sum_{i=1}^{n}\left\{\frac{D(y_i,\mu_i)}{\phi}\right\},$$

where D is the deviance function

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

The regression parameter estimates $\widehat{\boldsymbol{\beta}}$ are identical to those for the respective non-dispersed model and the overdispersion parameter ϕ is estimated by equating the Pearson X^2 statistic to the residual degrees of freedom (analogous to estimation of the residual variance from a normal model fit). For the overdispersed binomial model (2.3)

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

while for the overdispersed Poisson model (2.5)

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

McCullagh and Nelder (1989) discuss the advantage of basing estimation of ϕ on the Pearson X^2 statistic rather than using the residual deviance. The standard errors of the $\widehat{\boldsymbol{\beta}}$ will be as for the non-dispersed model inflated by $\sqrt{\widetilde{\phi}}$.

3.3 Extended quasi-likelihood

More complex overdispersion models can be generally described with a variance of the form $\mathrm{Var}(Y_i) = \phi_i(\boldsymbol{\gamma})V_i(\mu_i, \boldsymbol{\lambda})$, where both the scale parameter ϕ_i and the variance function $V_i(\cdot)$ may depend upon additional parameters. Nelder and Pregibon (1987) suggest estimating the unknown parameters in

the mean model ($\boldsymbol{\beta}$) and in the variance model ($\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$) by maximizing the extended quasi-likelihood (EQL) function

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

where D is the deviance function

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

Beta-binomial variance function

To illustrate this procedure we will consider the beta-binomial variance function 2.3 for proportion data. Here, we can take

$$V_i(t) = t(1-t/m_i)$$

and

$$\phi_i = 1 + (m_i - 1)\phi$$

giving a binomial deviance function, $D_B(y_i,\mu_i)$ and

$$Q^+ \;=\; -\tfrac{1}{2}\sum_{i=1}^{n}\left\{\frac{D_B(y_i,\mu_i)}{\phi_i} + log\left[2\pi\phi_i y_i\left(1-\frac{y_i}{m_i}\right)\right]\right\}$$

Differentiating to obtain the quasi-score equation for β_j gives

$$\frac{\partial Q^+}{\partial\beta_j} \;=\; \sum_{i=1}^{n}\frac{1}{\phi_i}\frac{y_i-\mu_i}{\mu_i(1-\frac{\mu_i}{m_i})}\frac{1}{g\prime(\mu_i)}x_{ij}$$

i.e. estimating equations for a weighted binomial model with weights $1/\phi_i = 1/[1+\phi(m_i-1)]$. These are simply the quasi-likelihood equations for a known value of the overdispersion parameter ϕ.

For the estimation of ϕ we need to solve

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

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^2_1 responses see McCullagh and Nelder (1989). We iterate between these two sets of estimating equations for $\boldsymbol{\beta}$ and ϕ until convergence, giving parameter estimates and standard errors, which are correct because of the asymptotic independence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\phi}$.

Brook's mixed strategy

Since the beta-binomial distribution is not in the linear exponential family, even for known values of ϕ, the maximum likelihood and quasi-likelihood estimating equations for $\boldsymbol{\beta}$ will be different. Brooks (1984) suggests a mixed strategy in which quasi-likelihood estimation for $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ 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 $(\boldsymbol{\beta}, \phi)$ 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 $\boldsymbol{\beta}$ with the beta-binomial variance function do not exactly maximize the beta-binomial likelihood.

Negative binomial variance function

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 $\boldsymbol{\beta}$, 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 (see Section 3.4) with estimating equation

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

3.4 Pseudo-likelihood

An alternative to extended quasi-likelihood is the pseudo-likelihood (PL) approach of Carroll and Ruppert (1988). Here estimates of $\boldsymbol{\beta}$ 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

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

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

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

where $r_i=(y_i-\mu_i)/\sqrt{V(\mu_i)}$, the unscaled generalized Pearson residuals, i.e. binomial type 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.

For the negative binomial variance function the estimating equation for k is

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

where $V_k(\mu_i) = \mu_i(1+\mu_i/k)$. Again, this can be set up as a gamma estimating equation using squared Poisson Pearson residuals as the y-variable, an identity link and a linear model with a fixed offset of 1 and the current estimate of μ_i as the explanatory variable. (Equivalently, we can use squared raw residuals and a model with μ_i as a fixed offset and μ_i^2 as the explanatory variable.)

3.5 Moment methods

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 $p = \dim(\boldsymbol{\beta})$, 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 $X^2 = \mathbf{E}[X^2]$, which gives simple one-step update for ϕ. On iterating this with quasi-likelihood estimation for $\boldsymbol{\beta}$ we obtain the same estimates as the degrees of freedom corrected moment method, since on convergence $\mathbf{E}[X^2] = n - p$.

For the negative binomial variance function 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 $\boldsymbol{\beta}$ 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,

$$\frac{d\log(1+\mu_i/k)}{dk} = -\frac{(\mu_i/k^2)}{(1+\mu_i/k)}$$

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.

Proportion data with equal sample sizes

For proportion data in which all of the sample sizes are equal to m, the beta-binomial variance function (2.5) reduces to constant overdispersion, as in (2.1). The weights in the quasi-likelihood estimating equations for $\boldsymbol{\beta}$ are all constant and so these equations reduce to those for the standard binomial model. The estimation of ϕ is also greatly simplified as

$$\frac{d\log(\phi_i)}{d\phi} = \frac{(m-1)}{[1+\phi(m-1)]}$$

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 $\mathbf{E}[D(y_i,\mu_i)] \neq \phi_i$. To take account of the estimation of $\boldsymbol{\beta}$ 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.

3.6 Non-parametric maximum likelihood

The basic idea behind this approach is to approximate the distribution of any unobserved random variables by a finite mass point distribution and use maximum likelihood methods to estimate the mixing distribution at the same time as any other unknown parameters. Aitkin (1995, 1996) has shown how this technique can be applied very easily to models with a random effect in the linear predictor, replacing 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, K say, and the weights, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_K$, and points, $u_1, \ldots, u_k$, for the quadrature become additional parameters in the model. The E-step now uses the current estimate of the mixing distribution in place of the quadrature points and weights. In the M-step the support points u_j are estimated by including a K-level factor in the model, while the associated probabilities are estimated from the conditional weights using

$$\widehat{\varepsilon}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n}.$$

Models are fitted for different values of K until the likelihood is maximized, giving a non-parametric maximum likelihood (NPML) estimate of the random effect distribution together with the regression parameter estimates. Rather surprisingly this typically happens with a fairly small value of K, so the technique is computationally feasible.

3.7 Bayesian approach

A Bayesian approach to overdispersion modelling is to take a two stage model and place priors on the parameters giving an hierarchical Bayesian model. For example for the logistic model with a normal random effect for overdispersion the basic model can be written as

$$Y_i \sim \text{Bin}(m_i, \pi_i), \quad \text{logit}(\pi_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + b_i, \quad b_i \sim N(0, \sigma^2)$$

and then independent "non-informative" priors can be specified for $\boldsymbol{\beta}$ and $\tau = 1/\sigma^2$. For models of this type Bayesian inference can be implemented

using Gibbs sampling or some other Markov chain Monte Carlo method. Gelman, Carlin, Stern, and Rubin (1995) gives a very good practical introduction to this approach to modelling, including some discussion of overdispersed data. Computation for these models can be done fairly easily in `BUGS`, see Spiegelhalter, Thomas, Best, and Gilks (1996). An advantage of this approach is that the computational method extends fairly easily to more complex random effect structures, see Gilks, Richardson, and Spiegelhalter (1996).

Chapter 4

Model selection and diagnostics

Selecting a model for a data set it is not an easy task. Sometimes the nature of the data tell us what is happening, but in general we need help from tests and plots to choose models which represent the data. Choosing between overdispersed models is still not completely understood, although several papers have appeared in recent years.

4.1 Model selection

4.1.1 Testing Overdispersion

Obtaining a goodness of fit test for overdispersed models is not as simple as in say fitting a binomial or Poisson model where the residual deviance or Pearson X^2 can often be used, at least as an approximate test. Because of the additional overdispersion parameters, which frequently act as some form of scale parameter, the situation has more similarity with fitting the normal distribution where the scale parameter σ is typically estimated from the residual deviance and thus the scaled deviance is just equal to the model degrees of freedom.

However, it is often possible to test the overdispersion model by comparison to the standard model fit. For a fully specified two stage model, testing

overdispersion frequently reduces to testing a specific value of a single additional parameter. For example, in comparing the negative binomial and Poisson distributions we can think of this as a test within the negative binomial family comparing $\widehat{\theta}$, the estimated value under the negative binomial, with $\theta = \infty$ which correspond to the Poisson model. The likelihood ratio test statistic compare the negative binomial and Poisson maximized log-likelihoods. Lawless (1987) notes that, since this involves testing a parameter value on the boundary of the parameter space, the appropriate asymptotic distribution for this statistic under the null hypothesis is one which has a probability mass of $\frac{1}{2}$ at 0 and a $\frac{1}{2}\chi^2_{(1)}$ distribution above 0.

Paul et al. (1989) discuss the distribution of the asymptotic likelihood test (LR) for goodness of fit of the binomial and multinomial distributions against the beta-binomial and Dirichlet-multinomial alternatives. They also derive the C(α) test (a generalization of Tarone's test) for goodness of fit of the multinomial distribution against the Dirichlet-multinomial alternative.

Dean (1992) derive tests for overdispersion with respect to a natural exponential family. He gives closed expressions of statistics (score tests) for testing constant overdispersion and two-stage models (beta-binomial, negative binomial and random effect models) against binomial or Poisson regression models.

Lambert and Roeder (1995) introduce a convexity plot, or C plot for short, that detects overdispersion and relative variance curves and relative variance tests that help to understand the nature of the overdispersion. They claim that convexity plots sometimes detect overdispersion better than score tests, and relative variance curves and tests sometimes distinguish the source of the overdispersion better than score tests.

4.1.2 Selecting covariates

Considering the constant overdispersion case if the fitted model is correct we have

$$\mathbf{E}(X^2) \approx (n-p)\phi$$

which suggests that now X^2 has a $\phi\chi^2_{n-p}$ distribution, where χ^2_{n-p} denotes a chi-squared random variable with $(n-p)$ degrees of freedom. Using the result that the deviance will usually be approximately equal to X^2, we might expect that in this situation the deviance will also have $\phi\chi^2_{n-p}$ distribution. This

result can be confirmed in McCullagh and Nelder (1989).

Once the overdispersion parameter has been estimated from fitting the full model, different sub-models can be fitted fixing its value and weigting the observations by $w_i = \frac{1}{\phi_i}$. Then two alternative (nested) models can be compared in the usual way. That is, the difference in the deviance for two models is compared with percentage points of the χ^2; a non-significant result means that the two models cannot be distinguished. This result is also used for other types of overdispersion models.

Fitzmaurice (1997) uses the Akaike information criterion (AIC) to compare deviances from different models.

4.2 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 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.

In the half-normal plot the ordered absolute values of some diagnostic quantity are plotted against half-normal scores (expected order statistics). For a sample of size n these are given by $\Phi^{-1}\{(i+n-\frac{1}{8})/(2n+\frac{1}{2})\}$. Departures from a straight line indicate non-normality and these plots can also be useful for outlier detection or checking the fit in a generalized linear model. However, because of the difficulty of deciding whether the plot differs significantly from a straight line, Atkinson (1985) suggests augmenting these displays with a simulation envelope. This also allows us to take account of any correlation structure in the diagnostics. The display is constructed as follows:

- Fit a model and calculate, $d_{(i)}$, the ordered absolute values of some diagnostic.
- Simulate 19 samples for the response variable using the fitted model and the same values for the explanatory variables.
- Refit the model to each sample and calculate the ordered absolute values of the diagnostic of interest, $d^*_{j(i)}$, $j = 1, \ldots, 19$, $i = 1, \ldots, n$.
- For each i, calculate the mean, minimum and maximum of the $d^*_{j(i)}$.
- Plot these values and the observed $d_{(i)}$ against the half-normal order statistics.

The minimum and maximum of the $d^*_{j(i)}$ provide an envelope which can be used to assess whether the observed values are consistent with the fitted model. If the fitted model is correct we would expect the plot of the observed values to lie within the boundaries of the envelope. Note that plotting against normal-based quantities is merely a convenience which allows us to assess the normality of the diagnostic quantity - in some situations this may not be appropriate and an alternative is to plot against the mean of the $d^*_{j(i)}$.

For generalized linear models these enhanced half-normal plots provide a useful tool for checking model assumptions. For overdispersed data the only change is that the simulation is now from the overdispersed model. If this corresponds to a completely specified distribution, such as beta-binomial and negative-binomial, this is perfectly straightforward. However, if the overdispersion model is only specified in terms of the mean and variance some ingenuity may be required to simulate data with the appropriate mean-variance

structure. For example, to simulate data from a binomial model with constant overdispersion we can rescale data simulated from a binomial model. A similar procedure can be used for the Poisson distribution. An alternative approach would be to simulate from any fully specified distribution with the correct mean-variance relationship. For the Poisson with constant overdispersion, ϕ, this could be done by simulating the Poisson parameter from a $\Gamma(r_i, \lambda)$ distribution, where $1 + 1/\lambda = \phi$.

Chapter 5

Examples

5.1 Binary data

5.1.1 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 Table 1.1. 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 5.1 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-

Table 5.1: 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*	EQL*	PL*	Moment*	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

*(df corrected)

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 $1 + \gamma(\overline{m} - 1)$, where the mean sample size is $\overline{m} = 39.6$, 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 2.6 are very similar to those using a beta-binomial form 2.5 – 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.

Table 5.2: Orobanche data: Deviances with overdispersion re-estimated for each model.

		Logistic	Beta-Binomial		
Source	NPML	Normal	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

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 2.2 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.

Table 5.3: Orobanche data: Parameter estimates for full interaction model.

Variable	Binomial	Williams Type II	Logistic Normal	NPML	Bayesian
Constant	$-.56 \pm .13$	$-.35 \pm .19$	$-.55 \pm .17$	$-.57$	$-.54 \pm .18$
Species	$.15 \pm .22$	$.07 \pm .31$	$.10 \pm .28$	.07	$.03 \pm .34$
Extract	$1.32 \pm .18$	$1.33 \pm .28$	$1.34 \pm .24$	1.30	$1.24 \pm .25$
Inter^n	$-.78 \pm .31$	$-.82 \pm .44$	$-.81 \pm .39$	$-.82 \pm .40$	$-.79 \pm .43$
σ		[.0249]	$.24 \pm .11$	.28	$.29 \pm .15$
				$-.31,\ .26$	
				$.45,\ .55$	

Using the NPML approach the mixing distribution is considered as a nuisance parameter and estimated for each model. Table 5.2 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, as in general are the parameter estimates, see Table 5.3. 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, see Figure 5.1 shows a strong monotonic relationship, indicating that the mixing distribution is picking up overdispersion.

It is interesting to note that the results from Tables 5.1 and 5.2, 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 5.2 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.

Some summary output from an hierarchical Bayesian analysis of this data is shown in Figure 5.2. In a qualitative sense the conclusions are very similar to those from the other approaches to overdispersion modelling.

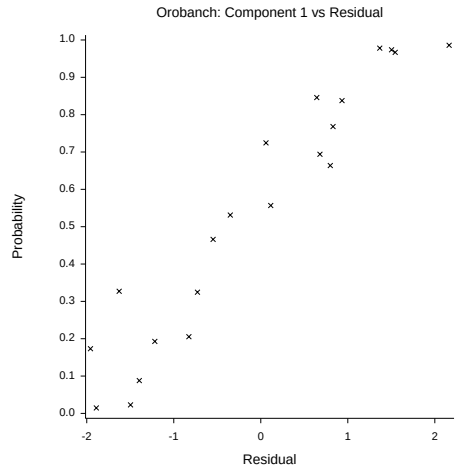

Figure 5.1: Orobanche Data. 2 component NPML fit, component probabilities

Figure 5.3 shows half-normal plots for some of the models considered in Table 5.3 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.

5.1.2 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 5.4 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 5.5 we present deviance contributions for the two factors in several models. The residual de-

Table 5.4: Trout egg data

Location in stream	Survival Period (weeks)			
	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

viance 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.

Table 5.5: Trout egg data: Deviances with overdispersion estimated from maximal model.

		Binomial		Constant	Beta-Binomial	
Source	d.f.	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

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 $1 + \gamma(\overline{m} - 1)$, where $\overline{m}$ 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.

The diagnostic plots discussed in Chapter 4 can also indicate the failure of various other model aspects. For example, Figure 5.4 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.

5.1.3 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 5.7 for several overdispersion models. With both constant overdispersion and a beta-binomial variance function the treatment effect is not significant.

5.1.4 Smoking and fecundability data

Weinberg & Gladen (*Biometrics*, 1986)
Retrospective study of pregnant women
Response: number of cycles to pregnancy

Table 5.6: Number of offspring surviving at 4 and 21 days by litter in 16 treated and 16 control rats.

	Control group Pups alive at			Treated group Pups alive at	
Litter	4 days	21 days	Litter	4 days	21 days
1	13	13	1	12	12
2	12	12	2	11	11
3	9	9	3	10	10
4	9	9	4	9	9
5	8	8	5	11	10
6	8	8	6	10	9
7	13	12	7	10	9
8	12	11	8	9	8
9	10	9	9	9	8
10	10	9	10	5	4
11	9	8	11	9	7
12	13	11	12	7	4
13	5	4	13	10	5
14	7	5	14	6	3
15	10	7	15	10	3
16	10	7	16	7	0

Cycle	Smokers	Non-smokers
1	29	198
2	16	107
3	17	55
4	4	38
6	9	22
7	4	7
8	5	9
9	1	5
10	1	3
11	1	6
12	3	6
> 12	7	2
Total	100	498

Table 5.7: Rat Survival data: Models with common and distinct overdispersion parameters – deviances and parameter estimates.

		Common Overdispersion				
		Binomial	Constant	Beta-Binomial	Logistic-Normal	Beta-Binomial
Source	d.f.	ML	QL	Moment	Moment	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	–

Geometric Model

Assume conception rate, θ, is constant over cycles.

$$P(\text{conception at } i\text{th cycle}) = (1-\theta)^{i-1}\theta,\ i = 1, 2, \ldots$$

or

$$\begin{aligned} P(X = i) &= (1-\theta)^{i-1}\theta \\ P(X > i) &= (1-\theta)^{i} \end{aligned}$$

Consider a single group, say smokers, with observations $n_1, n_2, \ldots, n_{12}, n_{12+}$

$$\begin{aligned} L(\theta|\mathrm{n}) &= \theta_{n_1}\{(1-\theta)\theta\}^{n_2}\cdots\{(1-\theta)^{11}\theta\}^{n_{12}}\{(1-\theta)^{12}\}^{n_{12+}} \\ &= \theta^{n_1}(1-\theta)^{N-n_1}\cdot\theta^{n_2}(1-\theta)^{N-n_1-n_2} \\ &\qquad \cdots\theta^{n_{12}}(1-\theta)^{N-n_1-n_2-\ldots-n_{11}} \\ &\propto \prod_{i=1}^{12} L_{B_i}(\theta|n_i) \end{aligned}$$

where $L_{B_i}(\theta|n_i)$ is the likelihood for an observed response n_1 from Bin$(N - \sum_{j=1}^{i-1} n_j, \theta)$ distribution.
Note: Each term corresponds to a binomial for the no. of conceptions at cycle i from the no. currently at risk.

cycle	1	2	3	4	5	...	11	12
conceptions	29	16	17	4	3	...	1	3
No at risk	100	71	55	38	34	...	11	10

Fitting using binomial with reciprocal link

$$\mathbf{E}[X] = 1/\theta$$

	Deviance	d.f.
`Constant`	83.92	23
`+smoking`	65.66	22

Smoking effect estimate = 1.448 (0.4224)

- constant conception rate not plausible
- suppose heterogeneity between couples, but constant conception rate for individuals over cycles.
  $\Longrightarrow$ declining rate over cycles
  sorting effect of heterogeneous population – those "still trying" have lower chance of success.

Beta-Geometric Model

Assume individual θs follow some distribution, e.g. θ ~Beta with mean μ, shape γ. Now,

$$\mathbf{E}[X - i | X > i] = \frac{1-\gamma}{\mu-\gamma} + \frac{i\gamma}{\mu-\gamma}$$

i.e. linearly increasing in over cycle.

- $\gamma = 0 \rightarrow$ no heterogeneity $\rightarrow$ Geometric
- likelihood again takes product binomial form where θ now depends on the cycle with
  $$\theta_i = c + d(i-1)$$
  c and d can be allowed to depend on covariates, e.g. smoking.

Beta-Geometric Fit

Model	Deviance	d.f.
`cycle`	39.67	22
`+smoking`	26.40	21
`+cycle.smoking`	26.48	20

Main effects model

parameter	estimate	s.e.
`1`	2.45	0.12
`cycle`	0.33	0.07
`smoking`	1.18	0.40

5.2 Count data

5.2.1 Pump failure data

Gaver and O'Muircheartaigh (1987) present data on the numbers of failures s_i and the period of operation t_i (measured in 1,000s of hours) ffor 10 pumps from a nuclear plant. The pumps were operated in two different modes; four being run continuously (C) and the others kept on standby (S) and only run intermittently.

s_i	5	1	5	14	3
t_i	94.320	15.720	62.880	125.760	5.240
mode	C	S	C	C	S
s_i	19	1	1	4	22
t_i	31.440	1.048	1.048	2.096	10.480
mode	C	S	S	S	S

Gaver and O'Muircheartaigh (1987) present data on the numbers of failures and the period of operation, t_i, (measured in 1,000s 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 5.8 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 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.

Table 5.8: Pump data: Deviances and parameter estimates

		Poisson	Constant	Negative Binomial		
Source	d.f.	ML	QL	ML	EQL	PL
			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

5.2.2 Fabric Fault Data

The data given in Table 5.9 are counts of the number of faults in rolls of fabric of different lengths. Figure 5.5 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

Table 5.9: Fabric data.

length of roll	faults	length of roll	faults
551	6	543	8
651	4	842	9
832	17	905	23
375	9	542	9
715	14	522	6
868	8	122	1
271	5	657	9
630	7	170	4
491	7	738	9
372	7	371	14
645	6	735	17
441	8	749	10
895	28	495	7
458	4	716	3
642	10	952	9
492	4	417	2

64.5 on 30 df, indicating overdispersion. Table 5.10 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.

Table 5.10: 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	

5.2.3 Quine's data

Aitkin et al. (1989) describe a data set on absence from school from a Sociological study of Australian Aboriginal and white children,see Table 5.11. The response variable of interest is the number of `days` absent from school cross-classified by age (`A`, 4 levels), sex (`S`, 2 levels), cultural group (`C`, 2 levels) and learning rate (`L`, 2 levels).

Using a Poisson model the residual deviance is very large even for the maximal model, showing strong evidence of overdispersion. A half-normal plot for the deviance residuals confirms this with the observed residuals almost completely outside of the simulation envelope, see Figure 5.6. The negative binomial distribution provides a possible overdispersion model for this data. Using the macros from Hinde (1996) we can fit a negative binomial distribution and use the half-normal plot to show that the data is consistent with this model, see Figure 5.6.

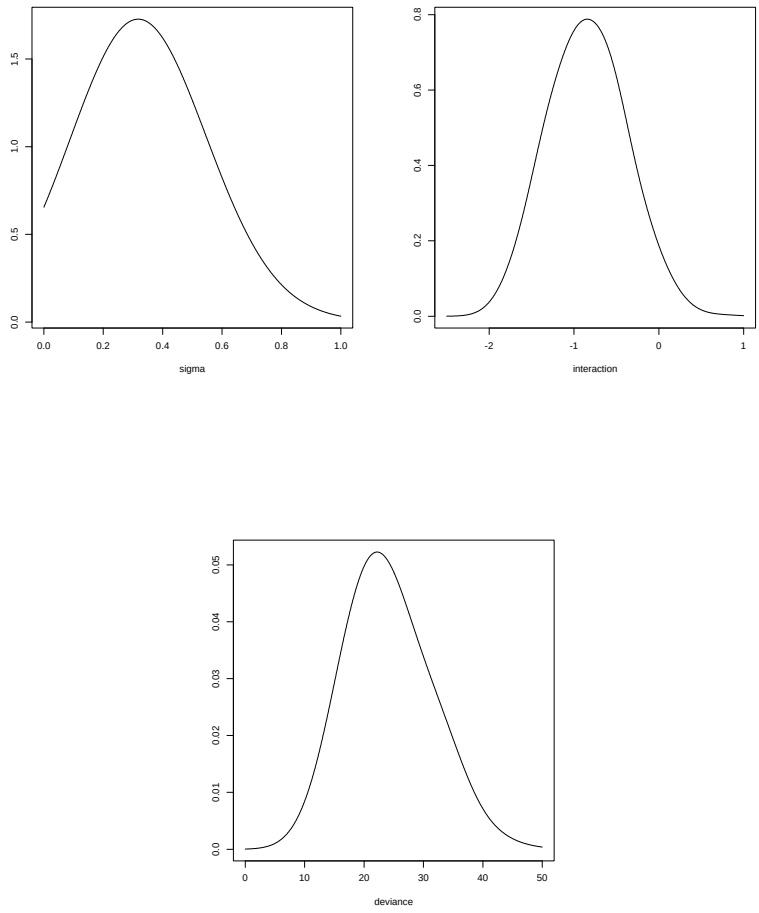

Figure 5.2: Orobanche Data. Hierarchical Bayesian analysis. Smoothed posterior distributions

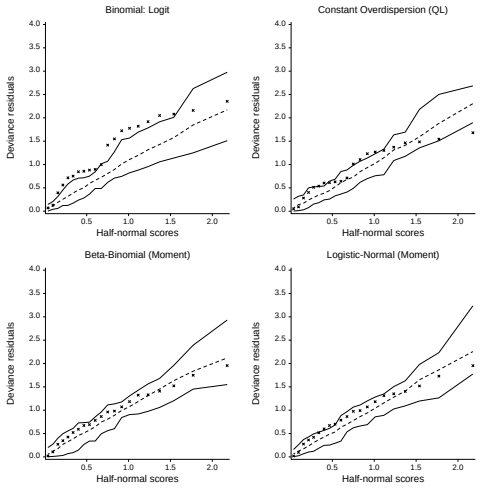

Figure 5.3: Orobanche Data. Half-normal plots: × – data; — simulated envelope

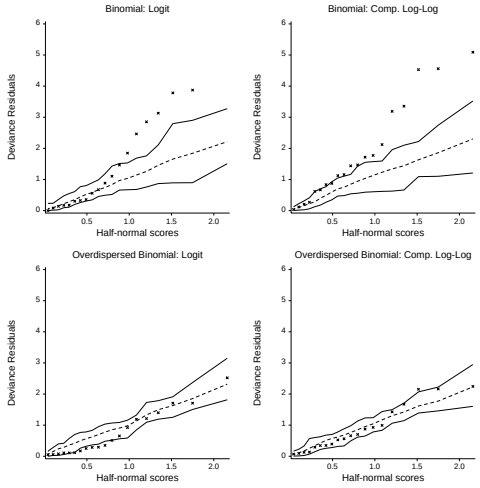

Figure 5.4: Manly Data. Half-normal plots: × – data; — simulated envelope

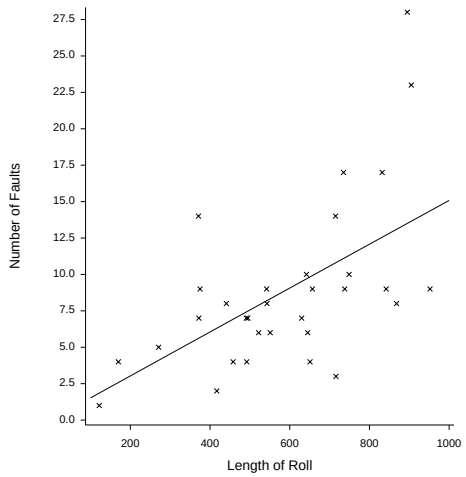

Figure 5.5: Fabric Fault Data. × – data; — Poisson fit

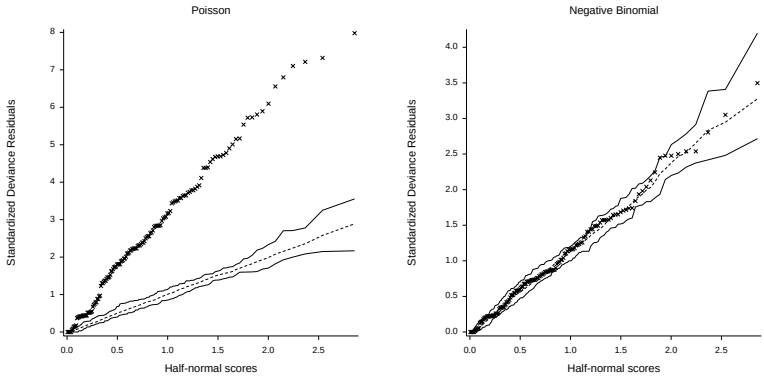

Figure 5.6: Quine data. Half-normal plots: × – data; — simulated envelope

Table 5.11: Quine's School absence data.

Days	C	S	A	L	Days	C	S	A	L	Days	C	S	A	L	Days	C	S	A	L
2	1	1	1	1	11	1	1	1	1	14	1	1	1	1	5	1	1	1	2
5	1	1	1	2	13	1	1	1	2	20	1	1	1	2	22	1	1	1	2
6	1	1	2	1	6	1	1	2	1	15	1	1	2	1	7	1	1	2	2
14	1	1	2	2	6	1	1	3	1	32	1	1	3	1	53	1	1	3	1
57	1	1	3	1	14	1	1	3	2	16	1	1	3	2	16	1	1	3	2
17	1	1	3	2	40	1	1	3	2	43	1	1	3	2	46	1	1	3	2
8	1	1	4	2	23	1	1	4	2	23	1	1	4	2	28	1	1	4	2
34	1	1	4	2	36	1	1	4	2	38	1	1	4	2	3	1	2	1	1
5	1	2	1	2	11	1	2	1	2	24	1	2	1	2	45	1	2	1	2
5	1	2	2	1	6	1	2	2	1	6	1	2	2	1	9	1	2	2	1
13	1	2	2	1	23	1	2	2	1	25	1	2	2	1	32	1	2	2	1
53	1	2	2	1	54	1	2	2	1	5	1	2	2	1	5	1	2	2	2
11	1	2	2	2	17	1	2	2	2	19	1	2	2	2	8	1	2	3	1
13	1	2	3	1	14	1	2	3	1	20	1	2	3	1	47	1	2	3	1
48	1	2	3	1	60	1	2	3	1	81	1	2	3	1	2	1	2	3	2
0	1	2	4	2	2	1	2	4	2	3	1	2	4	2	5	1	2	4	2
10	1	2	4	2	14	1	2	4	2	21	1	2	4	2	36	1	2	4	2
40	1	2	4	2	6	2	1	1	1	17	2	1	1	1	67	2	1	1	1
0	2	1	1	2	0	2	1	1	2	2	2	1	1	2	7	2	1	1	2
11	2	1	1	2	12	2	1	1	2	0	2	1	2	1	0	2	1	2	1
5	2	1	2	1	5	2	1	2	1	5	2	1	2	1	11	2	1	2	1
17	2	1	2	1	3	2	1	2	2	4	2	1	2	2	22	2	1	3	1
30	2	1	3	1	36	2	1	3	1	0	2	1	3	2	1	2	1	3	2
5	2	1	3	2	7	2	1	3	2	8	2	1	3	2	16	2	1	3	2
27	2	1	3	2	0	2	1	4	2	10	2	1	4	2	14	2	1	4	2
27	2	1	4	2	30	2	1	4	2	41	2	1	4	2	69	2	1	4	2
25	2	2	1	1	10	2	2	1	2	11	2	2	1	2	20	2	2	1	2
33	2	2	1	2	0	2	2	2	1	1	2	2	2	1	5	2	2	2	1
5	2	2	2	1	5	2	2	2	1	5	2	2	2	1	5	2	2	2	1
7	2	2	2	1	7	2	2	2	1	11	2	2	2	1	15	2	2	2	1
5	2	2	2	2	6	2	2	2	2	6	2	2	2	2	7	2	2	2	2
14	2	2	2	2	28	2	2	2	2	0	2	2	3	1	2	2	2	3	1
2	2	2	3	1	3	2	2	3	1	5	2	2	3	1	8	2	2	3	1
10	2	2	3	1	12	2	2	3	1	14	2	2	3	1	1	2	2	3	2
1	2	2	4	2	3	2	2	4	2	3	2	2	4	2	5	2	2	4	2
9	2	2	4	2	15	2	2	4	2	18	2	2	4	2	22	2	2	4	2
22	2	2	4	2	37	2	2	4	2										

Chapter 6

Extended overdispersion models

6.1 Random effect 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

$$h(\phi_i) = \boldsymbol{\gamma}^T \mathbf{z}_i$$

for some suitable link function h, usually the identity or the log. The vector of explanatory variables $\mathbf{z}_i$ 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 $\boldsymbol{\gamma}$ as in Sections 3.3 and 3.4; see McCullagh and Nelder (1989), chapter 12, for a detailed discussion of this.

6.2 Double exponential family

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, and when $\theta < 1$ there is overdispersion. For example, for overdispersed binomial data, the related double exponential family distribution can be written as

$$f(y_i;\pi_i,\theta_i) = c(\pi_i,\theta_i)\binom{m_i}{y_i}\frac{y_i^{y_i(1-\theta_i)}(m_i-y_i)^{(m_i-y_i)(1-\theta_i)}}{m_i^{m_i\theta_i}}\pi_i^{y_i\theta_i}(1-\pi_i)^{(m_i-y_i)\theta_i}$$

where $c(\cdot)$ is a normalising constant that is generally very close to 1. In this model the variance is effectively inflated by $1/\theta_i$ and the dispersion parameters θ_i can again be modelled by explanatory variables. 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).

6.3 Generalized linear mixed models

Another natural way to extend the category of two-stage models is through the addition of more complex random effects structures in the linear predictor, taking

$$\eta_i = \boldsymbol{\beta}^T\mathbf{x}_i + \boldsymbol{\gamma}^T\mathbf{z}_i$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{\gamma}$ is a vector of random effects and $\mathbf{x}_i$ and $\mathbf{z}_i$ 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).

6.3.1 Examples

Rat Survival Data

Returning to the rat survival data discussed in Section 5.1.3, examination of the data shows that there are different degrees of overdispersion in the two groups. 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 with a deviance difference of 5.77 on 1 df. A similar conclusion is obtained using a logistic-normal variance function (2.6), 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.

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.

Appendix A

Exercises

A.1 Melon organogenesis

The data presented in Table A.1 are from a 2x5 factorial, completely randomized tissue culture experiment. Small cuts of cotyledon (part of seed) of two melon varieties were grown in 4 different concentrations of BAP(mg/l). The purpose of this experiment was to see how organogenesis is affected by variety and concentration of BAP.

Table A.1: Melon organogenesis - Number of explants (y_i) regenerated out of 8 explants.

	Eldorado				AF-522			
Replicates	0.0	0.1	0.5	1.0	0.0	0.1	0.5	1.0
1	0	0	7	8	0	0	4	7
2	0	2	8	8	0	2	7	8
3	0	0	8	8	0	0	7	8
4	0	1	5	8	0	1	8	8
5	0	0	7	5	0	1	8	7

A.2 Apple tissue culture

The data presented in Table A.2 are from a completely randomized 2x5 factorial tissue culture experiment in which the experimental unit is a transparent plastic dish that is partitioned into a 5x5 array of individual compartments. A measured quantity of culture medium is poured into each compartment and a small piece of plant material, termed an **explant**, is placed on the surface of the medium. The dish is then kept in an incubator for several weeks, during which time new shoots may grow from the original explants, a process known as **regeneration**. One of the aims of this kind of experiment is to know the effect of experimental treatments on the proportion of explants that regenerate Ridout and Demétrio (1992).

Table A.2: Number of explants(y_i) of apple that regenerated out of 25, considering 16 dishes.

Dish	Explant type	Culture medium	Number regenerated (y_i)
1	A	X	8
2	A	X	10
3	A	Y	9
4	A	Y	11
5	B	X	9
6	B	X	10
7	B	Y	18
8	B	Y	12
9	C	X	7
10	C	Y	13
11	D	X	20
12	D	X	12
13	D	Y	20
14	D	Y	5
15	E	X	2
16	E	Y	13

A.3 Maize embryogenesis

The data presented in Table A.3 are from a completely randomized tissue culture experiment with three lines (L1, L2, L3) and three hybrids (H1=L1xL2, H2=L1xL3, H3=L2xL3) of maize. Imature embryos are inoculated and left in Petri dishes to form callus which can be embryogenic or not. The purpose of this experiment is to study the proportion of embryogenic callus.

Table A.3: Maize embryogenesis (table entries y_i/m_i).

After 45 days								
	Replicates							
L1	36/50	85/91	56/68	53/59	74/78	43/45	75/82	46/57
L2	82/102	75/103	65/76	79/86	89/110			
L3	113/127	86/87	102/106	41/44	79/108			
H1	46/50	65/71	57/67	73/86	57/70	73/80		
H2	81/93	118/127	93/107	80/89				
H3	100/127	69/79	92/99	92/95				
After one year								
	Replicates							
L1	1/45	3/76	0/53	4/44	5/58	3/25	4/62	0/37
L2	1/92	0/88	2/56	1/66	1/90			
L3	1/112	0/67	2/86	0/24	2/88			
H1	0/35	2/56	1/52	2/66	1/50	3/58		
H2	1/78	1/107	1/87	2/69				
H3	1/112	2/64	2/84	2/75				

A.4 Plum propogation

Table A.4. presents data from a 4x2 factorial plant propagation experiment in 5 randomized blocks conducted at East Malling International Research Station (Ridout 1990). The response variate is the number of cuttings which rooted out of a variable total number. The 'treatment' factor concern the stockplants from which the cuttings were taken. There were four types of stockplant: Normal, Large, Micro A and Micro B. Normal refers to standard plants of the species (a dwarfing plum rootstock), Large refers to a variant with a more vigorous growth habitat. Both Nomal and Large stockplants were themselves produced from cuttings (conventional propagation). Micro A and Micro B refer to stockplants which were produced by micropropation several years earlier. The A and B denote two different microprogation lines. The other treatment factor is the severity of prunning (Hard or Light) which

was applied to the stockplants. Hard prunning generally increases the proportion of cuttings which will root and the purpose of this experiment was to examine possible effects on rooting of different types of stockplant and to see to what extent these effects were dependent on the severity of prunning. The biological mechanisms underlying the enhanced rooting of cuttings from micropropagated plants are not completely understood, but it seems likely that the effect is one of 'rejuvenation'. Micropropagated plants exhibit typical juvenile growth characteristics, such as abundant production of lateral spines.

Table A.4: Plum propagation data (table entries y_i/m_i).

	Normal		Large		Micro A		Micro B	
Bock	Hard	Light	Hard	Light	Hard	Light	Hard	Light
1	7/17	5/22	10/37	7/30	7/24	12/27	15/21	13/37
2	9/17	0/16	5/20	12/26	12/13	10/21	15/23	10/17
3	10/12	7/13	4/20	10/19	11/18	13/21	11/21	16/23
4	13/21	7/15	19/21	6/30	14/18	15/22	21/24	15/26
5	10/21	0/17	10/27	1/14	21/26	22/39	15/25	11/30

Bibliography

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

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–262.

Aitkin, M., D. Anderson, B. Francis, and J. Hinde (1989). *Statistical Modelling in GLIM*. Oxford: Oxford University Press.

Aitkin, M. and B. Francis (1995). Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *GLIM Newsletter*, **25**, 37–45.

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

Atkinson, A. (1985). *Plots, transformations and regression*. Oxford: Clarendon Press.

Breslow, N. (1984). Extra-poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Breslow, N. and D. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brooks, R. (1984). Approximate likelihood ratio tests in the analysis of beta-binomial data. *Applied Statistics*, **33**, 285–9.

Carroll, R. and D. Ruppert (1988). *Transformation and Weighting in Regression*. London: Chapman & Hall.

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

Collett, D. (1991). *Modelling binary data*. London: Chapman and Hall.

Crowder, M. (1978). Beta-binomial anova for proportions. *Applied Statistics*, **27**, 34–7.

Dean, C. (1992). Testing overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, **87**, 451–457.

Demétrio, C. and J. Hinde (1997). Half-normal plots and overdispersion. *GLIM Newsletter*, **27**, 19–26.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM-algortihm. *Journal of the Royal Statistical Society, B*, **39**, 1–38.

Dobson, A. (1990). *An Introduction to Generalized Linear Models*. London: Chapman & Hall.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**, 709–21.

Engel, B. and A. Keen (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1–22.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer-Verlag.

Finney, D. (1971). *Probit Analysis* (third ed.). London: Cambridge University Press.

Firth, D. (1991). Generalized linear models. In D. Hinkley, N. Reid, and E. Snell (Eds.), *Statistical Theory and Modelling*, pp. 55–82. Chapman & Hall.

Fitzmaurice, G. (1997). Model selection with overdispersed data. *The Statistician*, **46**, 81–91.

Francis, B., M. Green, and C. Payne (1993). *The GLIM System. Release 4 Manual*. Oxford: Oxford University Press.

Ganio, L. and D. Schafer (1992). Diagnostics for overdispersion. *Journal of the American Statistical Association*, **87**, 795–804.

Gaver, D. and I. O'Muircheartaigh (1987). Robust empirical bayes analysis of event rates. *Technometrics*, **29**, 1–15.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (1995). *Bayesian Data Analysis*. Chapman & Hall.

Gilks, W., S. Richardson, and D. Spiegelhalter (Eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Hinde, J. (1982). Compound poisson regression models. In R. Gilchrist (Ed.), *GLIM82*, pp. 109–121. New York: Springer-Verlag.

Hinde, J. (1996). Macros for fitting overdispersion models. *GLIM Newsletter*, **26**, 10–19.

Jorgensen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society, B*, **49**, 127–162.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. London: Chapman & Hall.

Lambert, D. and K. Roeder (1995). Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association*, **95**, 1225–1237.

Lawless, J. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, **15**, 209–225.

Lee, Y. and J. Nelder (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, B*, **58**, 619–678.

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

Lindsey, J. (1989). *The Analysis of Categorical Data using GLIM*. Berlin: Springer-Verlag.

Lindsey, J. (1995). *Modelling Frequency and Count Data*. Oxford: Oxford University Press.

Lindsey, J. (1997). *Applying Generalized Linear Models*. New York: Springer-Verlag.

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

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (second ed.). London: Chapman and Hall.

McCulloch, C. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330–335.

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

Nelder, J. (1985). Quasi-likelihood and glim. In *Generalized Linear Models*, pp. 120–127. Berlin: Springer-Verlag.

Nelder, J. and Y. Lee (1992). Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society, B*, **54**, 273–284.

Nelder, J. and D. Pregibon (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

Nelder, J. and R. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, **135**, 370–384.

Paul, S., K. Liang, and S. Self (1989). On testing departure from binomial and multinomial assmptions. *Biometrics*, **45**, 231–236.

Prentice, R. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement error. *Journal of the American Statistical Association*, **81**, 321–7.

Ridout, M. (1990). An example of overdispersed proportions. *unpublihed*, 1–9.

Ridout, M. and C. Demétrio (1992). Generalized linear models for positive count data. *Revista de Matemática e Estatística*, **10**, 139–148.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks (1996). BUGS: Bayesian inference using gibbs sampling, version 0.5. Technical report, available from MRC Biostatistics Unit, Cambridge.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models and the gauss-newton method. *Biometrika*, **61**, 439–47.

Weil, C. (1970). Selection of the valid number of sampling uints and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food Cosmet. Toxicol.*, **8**, 177–82.

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

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