Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Fisheries Research 70 (2004) 319–337

FISHERIES RESEARCH

www.elsevier.com/locate/fishres

GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research

W.N. Venables*, C.M. Dichmont

CSIRO, 233 Middle Street, Cleveland, Qld. 4163, Australia

Abstract

This paper provides an overview of the modelling process using generalized linear models (GLMs), generalized additive models (GAMs) and generalized linear mixed models (GLMMs), especially as they are applied within fisheries research. We describe the essential aspect of model interpretation and construction so as to achieve its correct application. We start with the simplest models and show the progression from GLMs to either GAMs or GLMMs. Although this is not a comprehensive review, we emphasise topics relevant to fisheries science such as transformation options, link functions, adding model flexibility through splines, and using random and fixed effects. We finish by discussing the various aspects of these models and their variants, and provide a view on their relative benefits to fisheries research.

Keywords: Generalized additive model; Generalized linear model; Generalized linear mixed model; Spline regression; Statistical modelling strategies

1. Introduction

The goal of a good model is to represent the process it attempts to describe in as wide a range of the stimulus variables as possible without over-specification. To fix ideas, let y be a quantitative response variable and assume for the moment that $x = (x_1, x_2, \ldots, x_p)$ is a set of (initially) quantitative stimulus variables driving the distribution of y. Suppose z is a standard normal random variable which captures a degree of stochastic behaviour. This is a model that says almost nothing about the process other than that it has the form:

$$y = f(\mathbf{x}, z)$$

where f is a function yet to be determined. We are going to regard the random variable z as just another variable, centred on zero. Suppose we have a particular point, x_0, where we would like the model to perform well (and in some local region about it). Assuming the system is reasonably continuous and slowly varying, a natural way of proceeding is to use a first order Taylor approximation about the point $(x_0, 0)$ as a local approximation. Re-writing the derivatives at the central point as

* Corresponding author. Tel.: +61 7 3826 7251; fax: +61 7 3826 7304.
E-mail address: bill.venables@csiro.au (W.N. Venables).

0165-7836/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.fishres.2004.08.011

coefficients in a familiar notation, this gives:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j (x_j - x_{j0}) + \sigma z$$

which is a familiar linear regression model. If we want the model to capture the behaviour in a larger region about the central point, then it is natural to consider extending the Taylor series to second order, giving an expansion of the form:

$$y = \left\{ \beta_0 + \sum_{j=1}^{p} \beta_j (x_j - x_{j0}) + \sum_{k=1}^{p} \sum_{j=1}^{k} \beta_{jk} (x_j - x_{j0})(x_k - x_{k0}) \right\} + \left\{ \sigma + \sum_{j=1}^{p} \gamma_j (x_j - x_{j0}) \right\} z + \{\delta z^2\}$$

That is, we might go to a second-order polynomial model with powers and linear × linear interactions. The last two bracketed terms suggest, respectively, that we might also expect that variance heterogeneity and non-normality (mainly skewness and kurtosis) could start to play an increasingly important role as we require our model to apply in wider ranges. These are features of real modelling situations.

An important, if simple, message to take from this mathematical view is that most regression models are strongly empirical and should only be expected to apply in a limited region about some central point in the design (or *x*-variable) space. This should explain why, even when a regression line should logically go through the origin, it may be a better policy not to constrain it to do so if the origin is well outside the region in design space where observations are available. In the fisheries effort-standardization context then, we may be willing to relax some logical constraints or boundary conditions on the coefficients that we are estimating if the data we have are far from the boundary and, if by doing so, we improve the performance of the standardization where it really matters, namely near the data itself.

Generalized linear models (GLMs) attempt to accommodate variance heterogeneity and asymmetric, non-normal behaviour by offering a range of distributional types that cover at least the more common mean–variance relationships. GLMs are also useful for obviously non-normal data, such as binary data. The log-likelihood surface must be reasonably quadratic in a region about its maximum point for most parametric inferential methods, such as (model based) standard errors, confidence intervals and likelihood ratio tests, to be reliable.

One can allow for curvature in the regression surface, as above, by including polynomial terms. There are other, often better ways, of gaining flexibility in the regression surface such as using regression splines. We discuss this issue of flexibility in Section 4 first because it is important in itself and secondly because it provides a natural introduction to generalized additive models (GAMs).

Increasing the complexity of a regression model by including additional terms will increase the accuracy of the regression for the training data (the data used to estimate the values for the parameters of the model), but will also tend to decrease the accuracy of the model when it is used for prediction. This is because the extra complexity in the fitted model may actually be reproducing randomness.[1] This increased complexity can also affect the reliability of interpretations of the fitted model. The choice of the degree of complexity, then, has to balance accuracy in the training data with predictive accuracy or interpretative reliability. GAMs represent an extension to GLMs that partially automates this choice. Local smoothers, including smoothing splines, may be included in the regression function but estimation is not by maximum likelihood. Rather, a penalty term, which reflects the degree of smoothness in the regression, is added to the log-likelihood and this sum of terms is maximized. The relative weight given to log-likelihood and penalty is usually determined by cross-validation. We discuss GAMs in Section 5.

Generalized linear mixed models (GLMMs), represent a further and more fundamental extension of the initial regression model. In the general terms outlined above, they are best thought of as models where there are several independent places where a stochastic element enters the model:

[1] ‘Randomness is not the mere absence of pattern. Randomness can often show quite a definite pattern. The trouble is, next time it’s a completely different pattern’ (A.T. James, personal communication, 1975).

$$y = f(\mathbf{x}, z_1, z_2, \ldots)$$

where $z_1 \sim N(0, \sigma_1^2), z_2 \sim N(0, \sigma_2^2), \ldots,$ independently. The unknown variances are often the quantities of interest and are usually known as the *variance components*. GLMMs often arise where the parameters occur in natural groups. Instead of allowing each parameter in the group to count as a separate parameter, it may be natural to model them as being a sample from some distribution, typically normal. In this case, a group of parameters is replaced by a single variance component and the number of parameters is reduced. The price is we do not get, strictly speaking, an estimate of the individual parameters but rather predictors of them. In this respect they have a logical status more like residuals than parameter estimates.

There is a link between GLMMs and GAMs. The penalty imposed on the log-likelihood to ensure that GAMs remain economical with their use of parameters is analogous to the constraint imposed on the predictors in GLMMs, requiring them to behave like a sample from a specified distribution family. This often causes the predictors to be less volatile and less 'spread out' than would be separate parameter estimates, an effect known as 'shrinkage'.

One of the most important benefits of using mixed models is their capacity to 'borrow strength' from one part of the data to another, thus often providing a more realistic analysis of large fragmentary data sets, which are the norm in fisheries research.

2. Ordinary linear models

'Linear models form the core of classical statistics and are still the basis of much of statistical practice' (Venables and Ripley, 2002). This assertion is particularly true in quantitative fisheries research. Consider the following general set-up:

- There is a stochastic response variable of interest, say y.
- There are a number of candidate stimulus variables (or functions thereof), say $x_1, x_2, \ldots, x_p$ (which may be quantitative or qualitative).
- How the distribution of y depends on the fixed levels of the stimulus variables at which it is observed is described.

Note that we are using the generic x for any known function of the predictor variables, including the constant function 1 that occurs in the intercept term. If the stimulus variables are also stochastic, interest focuses on the conditional distribution of the response given the stimulus variables.

The linear model is perhaps the simplest and most direct approach to modelling this situation. The model is as follows:

$$y = \sum_{j=1}^{p} x_j \beta_j + \varepsilon = \eta + \varepsilon, \quad \text{where } \varepsilon \sim N(0, \sigma^2)$$

that is, the mean of the response depends linearly on the unknown coefficients β_j. The error term is then added to this linear function. The fixed part of this equation, represented here by η, is called the *linear predictor*. For our later purposes a slightly better way to write this model is:

$$y \sim N(\eta, \sigma^2), \quad \text{where } \eta = \sum_{j=1}^{p} x_j \beta_j$$

This emphasises that, in general, the error component is not simply 'added on' to the linear predictor, but is generated by the distribution of the response, in this case normal. Generalization will allow more distributions for the response and more general connections between the linear predictor and the mean of the response distribution.

We have the handy feature that the variable x_j is effective in influencing the distribution of y if, and only if, $\beta_j \neq 0$ because of the linearity in the unknown parameters. This important feature is preserved in all the generalisations of linear models we consider in this article.

The above simplistic model is unrealistic for many applications and the results may be misleading. Early attempts to accommodate stochastic behaviour in a response variable that is badly represented by a normal, homoscedastic, additive error term was to transform the response. This was so that the normal model was at least approximately true in the transformed scale. Although this can lead to problems in making inferences in the original scale, these can usually be overcome. We can represent this as follows:

$$t(y) \sim N(\eta, \sigma^2), \quad \text{where } \eta = \sum_{j=1}^{p} x_j\beta_j$$

where $t(\cdot)$ is a suitably chosen one-to-one transformation.

A different approach is to use generalized linear models, where some of the restrictive features of the simple linear model are relaxed. As we shall see, both approaches have their uses in practice.

3. Generalized linear models

The class of models known as generalized linear models, or GLMs, was formally introduced by Nelder and Wedderburn (1972), but the idea is much older. The techniques for fitting such models, for example, were essentially complete in Fisher (1954). The components of a GLM are as follows.

- The problem is again to model the distribution of a stochastic response variable, y, in terms of stimulus variables $x_1, x_2, \ldots, x_p$, or known mathematical functions of them.
- The distribution of y depends on the stimulus variables through a single *linear predictor:* $\eta = \sum_{j=1}^{p} x_j\beta_j$, where, in general, the x_j's are known functions of the stimulus variables, not necessarily simply the variables themselves.
- The mean of y is related to η by a known function called the *link* function:

  $$E[y] = \mu = \ell^{-1}(\eta), \quad \eta = \ell(\mu)$$

  Note that the link function transforms the mean into the linear predictor and not the other way round. Hence it acts in the same direction as a transformation of the response itself, from which the idea arose.
- The variance of y is a function of the mean: $\text{Var}[y] = \phi v(\mu)/A$ where ϕ is a possibly unknown, positive *scale parameter*, A is a known *prior weight*, and $v(\mu)$ is a known function of μ called the *variance function*.
- The distribution of y has a density of known form, namely

  $$f_Y(y;\mu,\phi) = \exp\left[\frac{A}{\phi}\{y\theta(\mu) - \gamma(\theta(\mu))\} + \tau\left(y, \frac{\phi}{A}\right)\right]$$

This distributional form can be shown to include the normal, gamma, Poisson and binomial distributions, as well as several others such as the beta, inverse Gaussian and negative binomial (if the extra variance parameter is known, see Section 3.7). Note that the relationship between the canonical parameter, θ, and the mean, μ, will depend on the particular distribution, and the relationship between μ and η is defined by the link function.

The theory of generalized linear models is concerned with a unified theory of estimation and testing. The standard reference is McCullagh and Nelder (1989), but there are many others (e.g. Chapter 7 of Venables and Ripley (2002)). GLMs have been used extensively in fisheries science. Their most common application is standardization of abundance indices based on commercial catch and effort data (e.g. Kimura, 1981; Punt et al., 2001; Maunder and Punt, 2004) or survey data (e.g. Stefánsson, 1996). However, applications have also included estimating selectivity of fishing gear (Myers and Hoenig, 1997), estimating bycatch catch rates (e.g. Ortiz et al., 2000; Ortiz and Arocha, 2004), estimating biological parameters such as growth (e.g. Bromley, 2000), and many others.

The choices for link functions, transformations (for error structure or zero data values) and model selection/complexity vary considerably in fisheries science, often for the same data types and problem. These issues are therefore discussed in further detail below.

3.1. The link function

The link function establishes the connection between the linear predictor, η, and the mean of the distribution, μ. There is a so-called 'natural link' for each distribution. The sense in which a link function is 'natural' is somewhat technical and such links are not necessarily very 'natural' in practice. Some sample information is lost if links other than the 'natural link' are used, but this is usually slight.

It is important to note that although the link function is in some senses similar to a transformation function, it only establishes a mathematical connection between parameters. A transformation function when applied to observations may be intended to simplify the connection between the mean and the response variables. It may also achieve other goals such as to stabilize the variance. See Section 3.2. Some special cases are:

(a) For the normal distribution, the natural link is the identity link, $\eta = \mu$, the variance function is constant, $v(\mu) = 1$, and the scale parameter is the variance, $\phi = \sigma^2$, which leads to ordinary linear models. These are sometimes artificially classified into regression, analysis of variance (ANOVA) or analysis of covariance (ANCOVA) models, based on old computational practices.

(b) In the case of binomial data, where the response is conventionally taken as the relative frequency, $y_i = s_i/a_i$ (where s is the number of successes and a the number of trials), the mean is a probability and hence must lie between 0 and 1. The linear predictor, on the other hand, is unbounded. Hence, the link function must map the real line into the closed interval [0, 1]. The natural link is the so-called logistic or logit link: $\eta = \log(\mu/(1-\mu))$, $\mu = e^{\eta}/(1+e^{\eta})$, but others are in common usage such as the probit link: $\eta = \Phi^{-1}(\mu)$, $\mu = \Phi(\eta)$, where Φ is the standard normal distribution function. The variance function has the form $v(\mu) = \mu(1-\mu)$ and the scale parameter is known, $\phi = 1$. The prior weight is the number of trials on which the observation is based, $A_i = a_i$. The difference between probit and logit links only becomes important if the probabilities being estimated are either very small or very close to unity, which typically require very large-sample sizes for effective inference to be possible. The two links generally give very close to equivalent results for intermediate probabilities.

(c) The natural link for the Poisson distribution is the log link: $\eta = \log(\mu)$, $\mu = e^{\eta}$, the variance function is $v(\mu) = \mu$ and, as in the case of the binomial distribution, the scale parameter is 1. Poisson models with log links are often called *log-linear models* and are used for frequency data. Often frequency data that does not strictly have a Poisson distribution can be analysed as if it had using ‘surrogate Poisson models’ (see Chapter 7 of Venables and Ripley, 2002).

(d) The gamma distribution has a natural link $\eta = 1/\mu$. The variance function is $v(\mu) = \mu^2$ and the scale parameter, ϕ, is generally unknown. The natural link is sometimes used in practice for the gamma distribution, but other links such as the log-link are more common. Note that natural link for the gamma distribution does not map the range of the mean into the unbounded natural range of the linear predictor. Therefore, the theory becomes only approximate in this case, though adequate for most applications. The exponential distribution is a special case of the gamma distribution.

3.2. Connection with transformation models

The classical method of dealing with non-identity connections between mean and linear predictor or non-constant connections between mean and variance has been to transform the response, i.e. the data are transformed using some function $t(y)$ prior to being analysed, so that some compromise between these two potentially conflicting requirements is met (e.g. Quinn, 1985; Richards and Schnute, 1992). Historically, the feature of having a consistent scale of variation has been (rightly) considered more important than achieving a simple connection between mean and linear predictor. Achieving a consistent scale of variation has therefore been given a degree of primacy when selecting a transformation.

The usual way of selecting a suitable transformation has been based on the assumption that, within the important region of variation of the random variable, the effect of a transformation can be captured adequately by a simple local linear approximation at the mean, i.e. if y has a distribution with mean μ and variance $\sigma^2(\mu)$, we want to find a transformation, $t(y)$ that makes the variance approximately constant. The linear approximation at the mean suggests that:

$$t(y) \approx t(\mu) + (y-\mu)t'(\mu)$$

Hence $E[t(y)] \approx t(\mu)$ and $\text{Var}[t(y)] \approx (t'(\mu))^2\sigma^2(\mu) = \text{const}$. This rough argument leads to the variance stabilising transformation given by:

$$t(x) \propto \int^{x} \frac{\mathrm{d}u}{\sigma(u)}$$

which suggests the square-root transformation for Poisson-like data, the arcsine-square root transformation for binomial-like data, and the log-transformation for data with approximately constant coefficient of variation: $\sigma(\mu) \propto \mu$.

Fishery data, for example catch rates, often have the property that the standard deviation increases with the mean approximately proportionally (e.g. Punt et al., 2000), that is the coefficient of variation is approximately constant. This may result from the fact that most factors that influence the model, including the random term, do so multiplicatively. Thus the natural model has the form:

$$y = \exp\left(\sum_{j=1}^{p} x_j\beta_j + \varepsilon\right), \quad \text{where } \varepsilon \sim N(0, \sigma^2)$$

A log transformation in this case exactly stabilises the variance and produces an identity link between mean and linear predictor in the transformed scale. Note that non-positive observations cannot occur in such a model, so if this is a feature of the data it must be captured by some additional feature of the model, or avoided by the unsatisfactory practice of adding a ‘small constant’ to all data before transforming (see Ortiz et al., 2000). It is incorrect to add an arbitrary value such as one to avoid the logarithm of zero. Further discussion of this issue can be found in Maunder and Punt (2004). By contrast, a normal model with log-link would involve a model of the form:

$$y = \exp\left(\sum_{j=1}^{p} x_j\beta_j\right) + \varepsilon, \quad \text{where } \varepsilon \sim N(0, \sigma^2)$$

corresponding to a non-linear regression with homoscedasticity in the original scale. Although the estimation of the mean parameters may be reasonable, the inferences based on this model depend on whether the variance is actually constant.

Often, as in the case of catch rate data, interest focuses on estimates of the mean in the original scale. Simply transforming back to the original scale produces estimates of the median rather than the mean, and since the lognormal distribution is massively right-skewed these are potentially highly negatively biased estimates of the mean. A simple correction is to add $\hat{\sigma}^2/2$ to the linear predictor before back-transforming, but this is also somewhat biased. Producing unbiased estimates of the mean of the lognormal distribution has received much attention in other contexts. One of the earliest solutions to this problem is given in Finney (1941) which produces the minimum variance unbiased estimate using an argument similar to the Blackwell–Rao theorem. Whether this degree of accuracy is needed in fishery applications is debatable, however, considering the usual roughness of the models used in the first place.

Another way of handling data with a multiplicative connection between mean and linear predictor and constant coefficient of variation is to use a distribution from the generalized linear family that has these properties, such as the gamma distribution with a log link. Alternatively, one could use a quasi-likelihood model (see Section 3.5) with log link and variance function $v(\mu) = \mu^2$. This approach has the advantage of working in the original scale and thus, in principle, avoids the problem of back transformation. This is an approach worth considering if inference in the original scale is paramount, but it is not entirely equivalent to the transformation approach.

Firth (1988) shows that even from an efficiency point of view the gamma model may have some minor advantages even if the lognormal model is the more appropriate. Wiens (1999) provides a simple example where the lognormal and gamma models lead to radically different outcomes, suggesting that the choice between these two models can be quite important. In our experience, the transformation approach is often more realistic for catch rate data, particularly since the gamma distribution has a much thinner upper tail than the lognormal. Very fat upper tails are often a feature of catch rate distributions. Another way of looking at this is to note that the error term also acts multiplicatively on the response for the transformation model. For the gamma model, the fixed factors do so, but the error term, which is not simply added to the linear predictor in this case, does not. In the transformed scale, diagnostics are certainly simpler and easier to appreciate.

3.3. Estimation and inference in GLMs

Estimates of the regression coefficients for normal linear models are obtained by least squares, and tests of significance are generally conducted by comparing the minimum sums of squares under different hypotheses using F-tests. Under the normal assumptions, these tests, or more generally inference methods, are ‘exact’, in the technical sense that no approximations are needed in their implementation. Generalized linear models offer a very natural extension of this situation in that:

- The computations involved in finding the maximum likelihood (ML) estimates of the regression parameters are very like those for the normal case, but must be applied iteratively to give successive approximations that converge to the ML estimates.

- The inference procedures use an analysis of deviance technique, essentially the likelihood ratio statistic, which essentially parallels the *F*-tests of normal theory linear models, and to which these procedures reduce in that case.

Generalized linear models therefore provide a uniform method of estimation and inference that is exact for the normal case with the identity link. Estimation is still exact maximum likelihood (ML) for the β parameters in other cases, but the inference methods are generally approximate, because the distribution theory behind analysis of deviance tests is based on the large-sample approximate distribution of the likelihood ratio statistic. Various alternative test procedures exist (e.g. the score test, Wald's test and the likelihood ratio test) which all coincide in the normal-identity case and are, in this sense, exact only in that case.

Chapter 7 of Venables and Ripley (2002) shows that the ML estimate of $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^{\mathrm{T}}$ can be found using the same computations as weighted regression applied iteratively. Given an initial estimate $\hat{\boldsymbol{\eta}}_0$, of the linear predictors (which may be essentially a link-transformed version of the observations, with some prudent modifications), initial estimates for the regression coefficients and variance weight function can be calculated. The weighted regression computation uses a constructed *working vector z* as the response and iterative weights given by the following formulae:

$$z_{0i} = \eta_{0i} + \frac{y_i - \mu_{0i}}{\mathrm{d}\mu_{0i}/\mathrm{d}\eta_{0i}}, \qquad w_{0i} = \frac{A_i}{v(\mu_{0i})}\left(\frac{\mathrm{d}\mu_{0i}}{\mathrm{d}\eta_{0i}}\right)^2$$

If X is the $n \times p$ design matrix and W_0 the $n \times n$ diagonal matrix of weights, then the next approximation to the β and linear predictor vectors are:

$$\hat{\boldsymbol{\beta}}_1 = (X^{\mathrm{T}} W_0 X)^{-1} X^{\mathrm{T}} W_0 z_0, \qquad \hat{\boldsymbol{\eta}}_1 = X\hat{\boldsymbol{\beta}}_1$$

from which iteration can usually proceed to convergence.

This iterative scheme attempts to find the maximum of the log-likelihood function, which, given ϕ, may be computed at any step as:

$$\log \hat{L}(\phi) = \sum\nolimits_{i=1}^{n} \left[\frac{A_i}{\phi}\{y_i\theta(\hat{\mu}_i) - \gamma(\theta(\hat{\mu}_i))\} + \tau\left(y_i, \frac{\phi}{A_i}\right)\right] \tag{1}$$

Notice that maximising this function with respect to the β parameters does not involve the second term (which is constant with respect to the β's). The ϕ parameter only occurs as a constant multiplier in the first term and hence the point at which the maximum occurs does not depend on ϕ. This is why the ML estimate of the β parameters may be found without knowledge of the scale parameter. (This important fact partly explains why the deviance is defined in the way that it is, as we discuss in Section 3.4.)

Eq. (1) is the profile likelihood for the scale parameter ϕ. In principle, the ML estimate of the scale parameter may then be found by maximising this quantity with respect to ϕ, although other estimators are often used, as we shall see.

The large-sample estimate of the variance matrix of the β parameters is then:

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \hat{\phi}(X^{\mathrm{T}}\hat{W}X)^{-1}$$

Tests on individual β coefficients using the standard test statistic:

$$z_i = \frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{\hat{\phi}(X^{\mathrm{T}}\hat{W}X)^{-1}_{ii}}}$$

as approximately standard normal under the null hypothesis are called Wald's tests. In fact, most of the large-sample inference procedures in generalized linear models can be deduced by using the analogy with weighted linear regression. In this very practical sense, generalized linear models offer a unified and very natural extension of linear least squares that is both computational and inferential.

3.4. *The deviance, its definition and its uses*

Many authors claim that the quantity called the deviance in generalized linear models is −2 times the maximum log-likelihood'. This is not strictly correct, and for some GLMs it is actually false and misleading.

To give a precise definition of the deviance for GLMs we need first to give a definition of a *saturated model*. This is a model with as many mean parameters as there are observations. In this case we may take the components of the mean vector as the parameters and it is easy to see that the ML estimates are the observations themselves: $\hat{\mu}_i = y_i$. We will denote the maximum of the log-likelihood function under the saturated model as:

$$\log \hat{L}_S(\phi) = \sum\nolimits_{i=1}^{n} \left[\frac{A_i}{\phi}\{y_i\theta(y_i) - \gamma(\theta(y_i))\} + \tau\left(y_i, \frac{\phi}{A_i}\right)\right] \qquad (2)$$

Notice that any model we may specify for the means, that is any design matrix X we may propose, specifies a model that is nested within the saturated model. To see this, note that that any model imposing a restriction on the mean vector through a real design matrix X *must* be a special case of the saturated model, which imposes no restrictions at all. The maximized log-likelihood for any real model, then, cannot exceed that for the saturated model.

Temporarily assuming ϕ is known (as it is for binomial and Poisson cases, but usually not otherwise) the likelihood ratio statistic for testing some specific model, say M, within the saturated model is then found by subtracting Eq. (1) from Eq. (2) and multiplying by 2:

$$\begin{aligned}\chi^2 &= 2(\log \hat{L}_S(\phi) - \log \hat{L}_M(\phi)) \\ &= \frac{1}{\phi}\sum_{i=1}^{n} A_i[\{y_i\theta(y_i) - \gamma(\theta(y_i))\} \\ &\quad - \{y_i\theta(\hat{\mu}_i) - \gamma(\theta(\hat{\mu}_i))\}] \overset{\text{def.}}{=} \frac{D_M}{\phi} \qquad (3)\end{aligned}$$

The quantity D_M so defined is the deviance for model M. Thus, the deviance may be defined as ‘the likelihood ratio statistic for testing any specific model within the saturated model, assuming the scale parameter is known and has the value 1’. The assumption is very important. The assumption is met for the binomial and Poisson distributions and the deviance is then merely a re-located version of $-2\log \hat{L}$, but for the normal and gamma distributions the assumption is usually not met and the deviance is not directly related to a likelihood ratio statistic at all.

The quantity inside the summation sign on the middle expression of Eq. (3) is called *the deviance increment*. These quantities in turn lead to the definition of *deviance residuals*, an important diagnostic tool to which we return below.

3.4.1. Distribution of the deviance; tests of fit

For the normal case, the expression for the deviance is the residual sum of squares:

$$D_M = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

and hence for the identity link this quantity does have a distribution proportional to the chi-squared: $D_M/\phi \sim \chi^2(n-r)$, exactly. This leads to the usual ‘variance component’ estimate of the scale parameter:

$$\tilde{\phi} = \frac{D_M}{n-p}$$

which for the normal case is the usual restricted maximum likelihood (REML) estimate of the variance, usually denoted by σ^2. REML estimation can be viewed as maximum likelihood estimation, but using a likelihood based on functions of the data which have a distribution depending only on the parameter of interest, in this case σ^2. In this sense, the likelihood is ‘restricted’, and the resulting estimate is usually closer to unbiased than the strict ML estimate, while retaining high efficiency.

The distributional properties of $\tilde{\phi}$ are virtually unknown for the gamma and inverse Gaussian cases. Nevertheless, if a fitted model for a distribution with ϕ known has $D_M/\phi \gg n-p$, the data is said to be ‘overdispersed’ with respect to the assumed distribution. Similarly if we have $D_M/\phi \gg n-p$ the data are ‘underdispersed’. This is less common, but can happen, for example, with binomial data when models have estimated probabilities close to 0 or 1. The variance component estimate of ϕ is very unreliable for some classes of data such as binary data or gamma data with values close to zero, and a different estimator is used based on the Pearson chi-squared statistic (see Section 4.4.5 of McCullagh and Nelder (1989)). In these cases, a view on whether the data are ‘overdispersed’ or ‘underdispersed’ relative to the assumed model should be based on this alternative estimator.

The assumption $\phi = 1$ holds for the binomial and Poisson cases, and the deviance is often used as a global ‘test of fit’, using the approximation $D_M \dot{\sim} \chi^2(n-p)$, with large values leading to rejection.[2] However, this approximate distribution has to be treated with some caution. It cannot be directly justified on the grounds of large-sample likelihood ratio theory. That theory pertains to testing one hypothesis within another, both of which have fixed degrees of freedom as the sample size increases. As the sample size increases the number of degrees of freedom associated with the saturated model by definition also increases, thus negating the assumption.

For the distribution of the deviance to be approximately chi-squared, a sufficient condition is that the distribution of the observations from which it is computed must become nearly normal. Thus, for the binomial case, we might then expect the distribution of the deviance, and hence the customary test of fit, to become approximately correct if the number of observations remains constant but the sizes of each of the numbers of trials, what we have called the a_i's above, increases. This is because by the central limit theorem each observation will become more nearly normally distributed. However, for the case of logistic regression with binary data, increasing the number of observations in general has an unknown effect on the distribution of the deviance. The approximate distribution theory will then usually not apply.

If a test of fit is required in the binomial or Poisson cases, a better proposal is to fit an enclosing model that includes all contemplated models as special cases and to test any given model within it. The degrees of freedom associated with the enclosing model should be relatively small compared with the sample size.

Likelihood ratio theory suggests that even if the deviance itself is not approximately chi-squared distributed, scaled differences of deviance (between fixed models with relatively low degrees of freedom compared to n) will have approximately chi-squared distributions. Thus if M and M_0 are two fixed models with p and p_0 degrees of freedom, and M_0 is nested within M, implying $p_0 < p < n$, and ϕ is known, then under reasonably general conditions:

$$\frac{D_{M_0} - D_M}{\phi} \mathrel{\dot{\sim}} \chi^2(p - p_0)$$

if M_0 is true. This provides the usual likelihood ratio test.

When ϕ is not known, the usual approximation uses an F-statistic. In this case the usual (but in some cases, somewhat speculative) approximation is:

$$\frac{(D_{M_0} - D_M)/(p - p_0)}{\tilde{\phi}} \mathrel{\dot{\sim}} F(p - p_0, n - p) \quad \text{if } M_0 \text{ is true}$$

again with large values leading to rejection. For the normal-identity case this is an exact result. For the normal case with non-identity links, the behaviour of this test statistic is not completely known. However, the approximation is usually assumed to be reasonably good, provided the model is not too non-linear. For other cases where ϕ is unknown, such as the gamma, the behaviour of this test statistic is not well known.

3.5. Quasi-likelihood

Generalized linear models offer considerable flexibility in modelling. The link function can be used to specify a non-linear connection between the linear predictor and the mean, and the distribution itself can be used to specify the variance function, that is, the connection between the variance and the mean. As we saw in Section 3.2, with transformations of the response itself, a single transformation had to be used to try to achieve both of these features, and the result was inevitably something of a compromise.

The estimation procedure and approximate inference methods presented above do not require the distribution to be stated explicitly, but rather rely on a number of functions to be specified. These are:

- the link function, $l(\cdot)$, which connects the mean to the linear predictor, and conversely;
- the variance function, $v(\mu)$, which specifies the mean–variance relationship up to proportionality;
- the deviance increment, which is only required at the inferential rather than the estimation stage.

This realisation led Wedderburn (1974) to introduce the notion of a quasi-likelihood model, which is only partially parametric in that it only requires these three ingredients to be specified rather than a fully parametric model.

Quasi-likelihood models (see Godambe and Heyde (1987) for a comprehensive treatment) can be shown to have various optimality properties regardless of the precise underlying distribution (e.g. Firth, 1987). Using quasi-likelihood models with the same link and

[2] The symbol $\mathrel{\dot{\sim}}$ is used to mean 'is approximately distributed as'.

variance function as the binomial or Poisson distributions produces the same estimates as for those distributions, but can be used with the scale parameter assumed unknown and estimated. These are sometimes called quasi-binomial or quasi-Poisson models. One way of allowing for overdispersion in inferences is to use a quasi-deviance to estimate a scale parameter. This device is closely related to one of the earliest ways of dealing with overdispersion (e.g. Finney, 1971).

Most software implementations of GLMs allow quasi-likelihood models to be specified in a straightforward way. For example, both S-PLUS and R allow a family argument in the GLM fitting function that may be used to specify a quasi-likelihood model in terms of the link and variance functions. R also has quasi-binomial and quasi-Poisson families that specify binomial- and Poisson-like 'distributions', but for which the scale parameter is assumed unknown and is estimated. The fact that no known discrete distribution has these properties is not an impediment to the nonparametric optimality properties still enjoyed by the estimation and inference procedures associated with quasi-likelihood methods.

3.6. Diagnostics and possible problems

Most of the diagnostic techniques for discovering problems with linear models can be applied fairly directly when using generalized linear models, with some caveats (e.g. Williams, 1987; Fox, 1997). The discovery of points of high leverage should use a projector matrix that takes into account the weights at the final stage of iteration, but otherwise the technique is identical to the ordinary linear regression case. This feature is automatically included in the facilities provided by the R software function 'influence.measures', to which readers may refer for additional examples (R Development Core Team, 2003).

Residual plots are also useful in most cases. Plots of sorted residuals against normal scores and against the individual predictor variables or the fitted values (or linear predictor values) are often used (e.g. Ortiz and Arocha, 2004) and should generally behave similarly to those for normal data. The residuals themselves are not uniquely defined and the S-PLUS and R software systems, for example, both allow four possible definitions of residual for generalized linear models. These all reduce to the same value for ordinary linear models, as might be expected.

Perhaps the most widely used definition of residuals for generalized linear models is the so-called 'deviance residuals' as introduced in McCullagh and Nelder (1989). The deviance residual for an observation is defined as the signed square-root of the deviance increment for that observation, where the sign is that of $y_i - \hat{\mu}_i$. Hence, just as the squares of the residuals in a linear model add to the residual sum of squares, the squares of the deviance residuals add to the deviance in a generalized linear model.

No definition of residuals is completely satisfactory for some classes of data. These include binary data and other frequency data with small numbers. In these cases, diagnostic investigations have to rely on other methods more specific to the particular problem (see, for example, Cox and Snell (1968) and Laird (1996)).

Discreteness in the data is not a particular problem for much of the machinery of inference using generalized linear models. Likelihood ratio tests, for example, rely on the likelihood being approximately quadratic in a sufficiently wide region about the maximum. Although there can be some problems with binomial models (e.g. the Hauck–Donner effect—Venables and Ripley (2002), p.197ff) this affects the convergence of the estimation process and Wald's tests more than likelihood ratio tests.

3.7. Overdisperson and model extensions

We have already discussed overdispersion as a potential problem with binomial- and Poisson-like data. An alternative to using quasi-likelihood models is to extend generalized linear models to incorporate extra components of variation which account for the increased dispersion. Technically this makes them generalized linear mixed models (GLMMs) of a kind, a subject to which we return below.

There is a considerable literature on this approach, with Williams (1982) one of the earliest papers on the binomial. One approach is to assume a two-stage model of the form:

$$y_i|\pi_i \sim B(n_i, \pi_i), \quad \pi_i \sim B(\delta, \gamma)$$

The marginal distribution of y_i is then beta-binomial, and the extra variation induced by the beta component (the second B above) induces the extra dispersion in the model relative to the binomial (the first B) alone. Williams (1982) also describes two other possible approaches.

Another distribution often used for frequency data overdispersed relative to the Poisson is the negative binomial distribution (e.g. Bannerot and Austin (1983) for catch rate data). This may also be specified using a two-stage formulation, namely:

$$y_i|z_i \sim \mathrm{Po}(\mu_i z_i), \quad z_i \sim \Gamma(\theta, \theta)$$

The mixing gamma distribution has $E[z_i] = 1$ and $\mathrm{Var}[z_i] = 1/\theta$ so the distribution of y_i approaches the Poisson again as $\theta \to \infty$, i.e. as the mixing variable becomes constant. It is also easy to see by conditional expectation and conditional variance arguments that:

$$E[y_i] = \mu_i, \qquad \mathrm{Var}[y_i] = \mu_i + \frac{\mu_i^2}{\theta}$$

The distribution has variance which is quadratically related to the mean rather than linearly as in the Poisson and quasi-Poisson cases for small values of θ. The marginal distribution of y_i is given by:

$$f(y_i; \mu_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta) y_i!} \frac{\theta^\theta \mu_i^{y_i}}{(\theta + \mu_i)^{\theta + y_i}}, \quad y_i = 0, 1, 2, \ldots$$

which, if θ is known, conforms to the generalized linear model distributional form. In waiting time data this occasionally is the case, but, in general, θ will be unknown. Thus, the model may be fitted by adapting the iterative scheme to accommodate the extra parameter (see Chapter 7 of Venables and Ripley (2002) for a more complete discussion, including examples).

4. Achieving flexibility in the linear predictor; moving to GAMs

We have noted above that including polynomial terms in the linear predictor is a natural way of enlarging the region within which an empirical regression relationship may be useful. This is analogous to increasing the number of terms in an approximating Taylor series to an unknown function. A problem with polynomials, in mathematical terms, is that the local behaviour determines the global behaviour. Often good behaviour in one region is bought at the expense of extremely poor behaviour elsewhere. Within the generalized linear modelling family, one way around this is to use a link function in the model that incorporates some non-linear behaviour. For example, the mean of a binomial proportion can only lie between 0 and 1; the link function ‘wraps’ the linear predictor into this finite range. Another way to extend the range of applicability of a model is to use a genuinely non-linear regression, which takes the model outside the generalized linear modelling family.

Remaining within the GLM family, this obstacle can sometimes be overcome by using a family of functions that can adapt to the local behaviour of the regression function almost independently in several regions at once. One such family is the spline family.

Spline functions (e.g. de Boor, 1978) are *piecewise* polynomials, usually over a finite range. At the ‘knots’, the points where the polynomial pieces join, the function is constrained to remain continuous and smooth. At the ends of the range (the ‘boundary knots’), a further constraint is applied to identify the function. For example, so-called ‘natural cubic splines’ are piecewise cubic polynomials with continuous first derivatives at the knots and constrained to be linear outside the boundary knots. The two important properties of splines from a data analysis point of view are.

- They may be expressed as linear combinations of known basis functions, analogous to the power terms used to define polynomials, and hence may be fitted with no more difficulty than polynomials.
- They are more flexible than, say, polynomials or harmonic functions. This is because being piecewise functions with discontinuous higher derivatives at the knots their local behaviour at a point does not entirely determine their global behaviour, i.e. they can ‘adapt’ to local conditions almost independently in several parts of their range.

One price that has to be paid for this less rigid behaviour relative to high-order polynomials is in the interpretability of the coefficients. Spline regressions are most easily appreciated graphically through the predicted or fitted values in the regression rather than through the values for their coefficients. In testing a spline model the entire block of terms should either be in the regression or out; in general it makes little

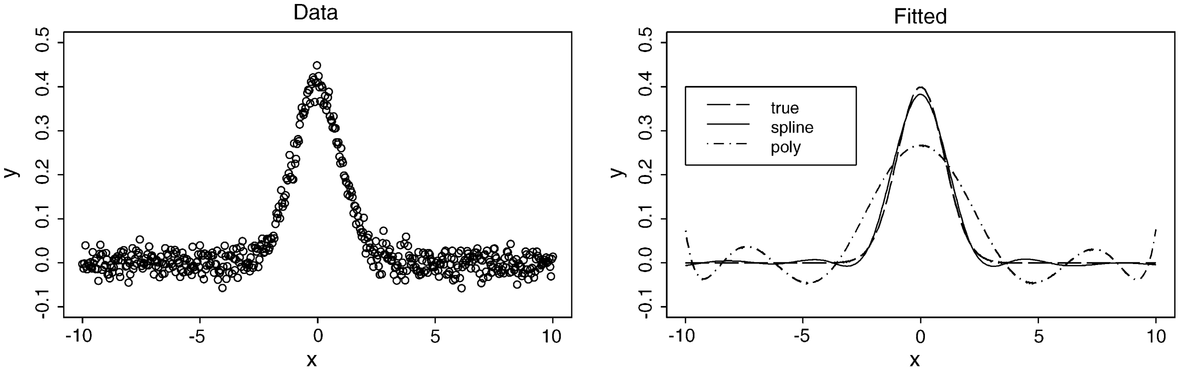

Fig. 1. An artificial example showing the true regression function (solid), the natural spline regression approximation with 10 degrees of freedom and knots at $x = -8, -6, \ldots, 6, 8$ and boundary knots at $x = -10, 10$ (dashed) and a polynomial regression, also with 10 degrees of freedom (dotted).

sense to consider the separate significance of individual terms.[3]

Fig. 1 shows an artificial example of a regression in one variable that is almost constant for much of its range but with a 'hump' in one region. The true function is a standard normal density over the region $(-10, 10)$ and there are 501 evenly spaced observations. The (artificial) data are shown in the top left panel and in the top right panel the true function is plotted together with the least squares estimates of a polynomial and a natural spline regression, both of which use 10 degrees of freedom. The polynomial fails to capture the behaviour of the regression virtually everywhere, whereas the natural spline does a reasonable job. In particular, the polynomial fails spectacularly at the ends of the range; the natural spline would not fare that much better for extrapolation very far from the ends of the range, but is certainly more stable. The only way we know of achieving stable extrapolation would be to fit a non-linear regression of the true form, namely $y = \exp(\beta_0 + \beta_1 x + \beta_1 x^2) + \varepsilon$, which has a mean that is a non-linear function of the unknown parameters. This is a non-linear regression, but the non-linearity can be captured by a log-link, thus remaining within the generalized linear model family. The fitted model and the true regression curve virtually coincide in this case, but it does require that the data analyst knows the true form of the regression.

One of the earliest papers on spline regressions (Boneva et al., 1970) suggested using them for exploratory purposes and this is still possibly their most effective use. Most statistical software platforms that provide generalized linear model fitting facilities now provide for spline regression, usually with a choice of 'natural' or 'B-spline' bases. The distinction between these lies in the identification constraint imposed at the ends of the range, but for most regression purposes the two bases are virtually equivalent.

4.1. Cross-product terms

When there are several predictor variables, it is common practice to fit independent spline terms in each, at least during exploratory analyses:

$$\eta = \beta_0 + s(x_1, \boldsymbol{\beta}_1) + s(x_2, \boldsymbol{\beta}_2) + \cdots + s(x_p, \boldsymbol{\beta}_p)$$

(where, despite notational appearances, all β's occur linearly). The individual terms are then easy to plot against the variable on which they depend, and usually easy to interpret. This presumes that the linear predictor can be so written, however, and that cross-product terms between different variables are not required. Unfortunately this is not necessarily the case in practice. The estimates of the 'main effects' (here spline) terms can be very misleading if important cross-product terms are omitted.

[3] To construct a nested sequence of natural cubic spline models in one variable of increasing complexity, the knot sequences have to be nested in the obvious sense. The complexity could be increased in this way and the testing theory would be entirely analogous to testing polynomial models of increasing degree. However, we are not aware of this being done commonly in practice.

Including cross-product terms greatly complicates the interpretation, and analysts are often loathe to include such terms for this reason. There is no easy solution to this, however, and the analyst has a responsibility to address this issue seriously.

One general way around this problem can be to choose variables in such a way that one would not expect, a priori, interactions among them to be very large. For example, it may be tempting to choose additive spline terms in latitude and longitude for models where graphical location is an important predictor. However, for a coastal fishery it may be more natural to take the distance along the coastline as one geographical co-ordinate and distance from the coastline out to sea as the complementary co-ordinate (Venables and Dichmont, 2004). If the GLM is describing, for example, fish abundance measures, it is easy to envisage situations where the latitude and longitude predictors are strongly interacting, but alternative geographical predictors are not.

5. Generalized additive models

Spline regression models can be parametrically very expensive and easily lead to over-fitting. Generalized additive models address this problem by deliberately fitting a model with a large number of parameters, but compensating for this by estimating them using a penalized likelihood, with the balance between likelihood and penalty chosen by cross-validation.

The use of GAMs in fisheries science is much less common that GLMs, but their use has increased substantially over the last decade. Many scientists are using GAMs instead of GLMs, and, as a result, the most common use in fisheries science is similar to that for GLMs, namely standardization of abundance data (e.g. Walsh and Kleiber, 2001). Most studies use a combination of commercial and/or survey data together with geographic and environmental variables for understanding and predicting abundance (e.g. Borchers et al., 1997; Bigelow et al., 1999; Denis et al., 2002; Brynjarsdóttir and Stefánsson, 2004), stock or species structure (e.g. Cardinale and Arrhenius, 2000; Venables and Dichmont, 2004) or distribution (e.g. Wright et al., 2000).

Rarely does one find that a mixture of GLMs and GAMs has been used in the same study. One example is, however, Agnew et al. (2003) where a combined GLM/GAM of parasite abundance in *Micromesistius australis* was modelled as a binomial GAM on the presence/absence component and infection intensity was modelled using a simple GLM.

A generalized additive model (Hastie and Tibshirani, 1990) is a generalized linear model that allows an extended form of linear predictor, namely:

$$\eta = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

where the $f_i(x_i)$ terms may well involve unknown parameters, but these are suppressed in the notation. The $f_i(x_i)$ terms are, in general, 'local smoothers', meaning they may be explicit functions or they may be, for example, 'loess' terms, which are more like prescriptions for achieving a local approximation by weighted averaging of near neighbours than explicit function definitions. One common choice is the so-called 'smoothing splines', which are splines with knots at each distinct value of the variable. If the estimation were not penalized, 'smoothing splines' would interpolate the data. The use of smoothing splines in regression is discussed comprehensively in Wahba (1990).

In addition to local smoother terms, generalized additive models may contain other terms with fixed degrees of freedom such as polynomials, harmonic terms or ordinary splines. These are omitted from the discussion here for simplicity. They enter the likelihood but not the penalty terms to be described below.

The comments on cross-product terms made in Section 5.1 still apply: it is assumed that the analyst has chosen the *x*-variables in such a way that cross-product terms are not likely to be important relative to the terms in each single variable that remains. If this is possible, it implies that the effect of each variable on the response is summarised by the $f_i(x_i)$ term which includes it. This makes interpretation of the model relatively easy.

5.1. *Estimation with penalties*

If L is the likelihood function (initially assuming the scale parameter has a known value of 1), estimation is achieved by minimising the penalized negative log-likelihood:

$$-\log L + \sum_{j=1}^{p} \lambda_j \int (f_j''(x_j))^2 \, \mathrm{d}x_j$$

where first term measures the closeness of the fit to the data and the second term measures the degree of ‘roughness’ in the regression function. The λ_j’s are ‘tuning’ constants that effect the trade-off between accuracy and smoothness, and are generally chosen by cross-validation. The fitting process is fully described in Hastie and Tibshirani (1990).

5.2. Discussion of GAMs

In our view, generalized additive models, if the problem of cross-product terms can be satisfactorily settled a priori, can be a powerful exploratory tool highlighting unexpected behaviour of some variables in their influence on the distribution of the response. However, they come at a relatively high cost. While the interpretation of the results may be relatively simple, at least graphically, any formal inference procedure, such as hypothesis tests or even obtaining confidence intervals for the fitted values, can be somewhat problematical. It is even possible for the deviance to increase in some cases if additional terms are added to the model, leading naive analysts to arrive at notional chi-squared test statistics that are negative. This apparently anomalous behaviour is resolved by noting that the fitting process does not minimize the deviance but rather the penalized deviance and the tuning constants may easily change considerably between any two models, implying a different trade-off.

In fisheries research, the added flexibility of generalized additive models over, for example, generalized linear models with fixed spline, or even polynomial terms may sometimes be necessary, but in our experience this is uncommon. We find that most applications in fisheries research are adequately handled by a judicious use of spline and polynomial terms (or harmonic terms if the function has a known period), and the stable and relatively straightforward inference procedures that this allows is a highly important bonus. This is not to say that generalized additive models might not be used for exploratory purposes prior to an analysis. In some cases, the extra flexibility of, say, smoothing splines with penalized estimation may really not be adequately replaced by fixed knot splines, but in this case we urge users to be cautious with the approximate inference procedures usually suggested.

6. Generalized linear mixed models

6.1. Mixed models

Before describing GLMMs we find it useful to present an artificially simple example based on ordinary mixed models. Consider perhaps the simplest of all possible linear models, that of a single mean:

$$y_i = \mu + \varepsilon_i$$

If we do not specify this model any further and regard all unknowns as parameters, there are more parameters than observations, namely $\mu, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$. If we identify the problem by imposing some constraint, for example, $\sum_{i=1}^{n} \varepsilon_i = 0$, then the model is saturated and the parameters are merely a different way of presenting the full data set and nothing is achieved. If, on the other hand, we extend the model in the usual way by requiring that $\varepsilon_i \sim N(0, \sigma^2)$, independently, then the number of parameters condenses to two: μ and σ^2, with estimates $\bar{y}$ and $s^2 = 1/n - 1\sum_{i=1}^{n}(y_i - \bar{y})^2$ and the status of the differences, or residuals $\hat{\varepsilon}_i = y_i - \bar{y}$ changes from parameter estimates to ‘predictors’ of the value of the unobserved variable. In this sense, all sensible models are ‘mixed’ models in that they have systematic and random components. It is more usual, however, to reserve this description for the situation where more than one random term enters the model.

Robinson (1991) provides a good reference on linear mixed models and best linear unbiased prediction.

6.2. GLMMs proper

Generalized linear mixed models are like generalized linear models but some of the terms in the linear predictor are random variates. The model may be formally described conditionally as (now using an obvious vector notation):

$$y|\boldsymbol{\zeta} \sim \mathrm{GLM}(\boldsymbol{\eta}, \phi),$$

$$\text{where} \quad \boldsymbol{\eta} = X\boldsymbol{\beta} + Z\boldsymbol{\zeta} \quad \text{and} \quad \boldsymbol{\zeta} \sim N(0, \Sigma(\theta))$$

Note that the design matrix is expressed in two parts, X for the fixed effects and Z for the random effects. The random effects, ζ, need not be multivariate normally distributed, but this is the common practice. Since the random effects are not observed, the true likelihood is based on the marginal distribution of y, which, in principle, can be obtained by integration. For a more extensive discussion of this possibility and its practical limitations, see Chapter 7 of Venables and Ripley (2002). The marginal density of y is:

$$f_Y(y; \boldsymbol{\beta}, \phi, \theta) = \int f_{Y|\zeta}(y|\zeta; \boldsymbol{\beta}, \phi) g_Z(\zeta; \boldsymbol{\theta})\, \mathrm{d}\zeta$$

where the integral over the multivariate distribution of the random effects is generally not tractable.

This rather formal definition can obscure both the simplicity of the method and its flexibility to capture real features of an actual situation. For example, random effects will usually be nested at different levels, such as ‘between areas’ and ‘between vessels within areas’, assuming areas can be modelled as random. For example, Lai and Helser (2004) model growth data where individuals are nested within survey strata.

Also, longitudinal data will normally capture ‘within vessel, between times’ variation either with an explicit correlation structure or by assuming an additive random vessel effect. In other cases, we may have random slope and intercept terms, which will normally be correlated, where the individual random regression lines occur, say, within vessel over time.

Estimation in GLMMs is still a research topic, but several approximate techniques are now (still somewhat cautiously) gaining acceptance. Diggle et al. (1994) and Laird (1996) summarise the theory and other issues related to GLMMs in the context of longitudinal data, a common context in which mixed models arise. Longitudinal data are cases in which several measurements are made on the same experimental units (e.g. vessels) over time. The fact that the same vessel is used normally induces correlations among the observations, which are important for the model to capture. One very effective way of achieving this is to attribute a random effect to each vessel, implying a GLMM. The use of GLMMs in fisheries science nevertheless remains fairly rare with only a few examples in the mainstream fisheries literature (e.g. Cooke, 1997; Squires and Kirkley, 1999; Olsen, 2002; Brandão et al., 2004).

Several closely related approximate procedures are now most often known as ‘penalized quasi-likelihood’ methods. Schall (1991) used the iterative weighted linear regression analogy, and suggested an estimation scheme for GLMMs that amounted to an iterative weighted version of fitting linear mixed effects models. Breslow and Clayton (1993) developed a similar method using the Laplace method for approximating the multiple integral involved. This latter method and their term, PQL, is now perhaps the most commonly used method, although others are gaining in popularity (see also Wolfinger and O’Connell (1993)). Some software is becoming available that uses numerical integration or Markov Chain Monte Carlo (MCMC) methods. See, for example, the GLMMGibbs and lme4 packages in R. For a very different approach, based on the EM algorithm, see van Dyk (2000) and references therein.

6.3. GLMMs: a generalized example

The implications of using GLMM are probably best conveyed by a concrete, but generalized, example. In many fisheries, the catch is a combination of several species. However, because of the substantially different biology of the species, separate stock assessments are needed for each species in the catch. Survey or observer data, even though very patchy in distribution and time, can be used to gain information on the relative species proportions in the catch in different areas and at different times of the season. We consider the problem of building a GLMM for predicting the proportion of the catch, by weight, of a species within a catch group. (A specific example is described in Venables and Dichmont (2004).)

The observations driving the model are the total weights, T, of catch in survey trawls, and the proportions of the weight for one of the species, y. Although the proportions are not binomially distributed, it is reasonable to consider a quasi-likelihood model that has a mean and variance function similar to the binomial. Hence, if μ is the true proportion, we propose a quasi-binomial model of the form:

$$y \sim \text{quasi-binomial}\left(\mu = \frac{e^{\eta}}{1+e^{\eta}}, \text{Var}[y] = \frac{\mu(1-\mu)}{T/\phi}\right)$$

(where here, ‘$\sim$’ means ‘is modelled as’). The variables available for the linear predictor include fine- and large-scale spatial coordinates, fine- and large-scale time, periodic time, and geographical variables such as depth and sediment type. In the specific example referred to above, the final model contained a mixture of spline terms, a linear term, four harmonic terms and an interaction term. There were also two random terms, which convert the model from a GLM into a GLMM. These were random increments for:

1. the season in which the survey took place, and
2. the stock region within each season.

These, respectively, allow for changes among seasons, and differences among stock regions within seasons, not otherwise captured by the model. They introduce two additional components of variation into the model. More importantly, they induce correlations among observations at two levels, namely within the same stock region and season, and within the same season. These may be important in allowing for unmeasurable factors influencing the proportions that need to be included in this surrogate way to enable the effect of other factors to be estimated accurately. If these random terms (technically a random main effect and a random interaction) were estimated as fixed effects, there would be two main differences:

1. the number of parameters in the model would be greatly increased, possibly leading to overparameterization and, more importantly,
2. the fixed effect model would not allow any future prediction, because, to make a prediction for a given future fishing season, we would need to know the unique increment for the season, as well as the unique increments of the stock areas within the season.

The random terms do not contribute to the fixed part of the mean, but the variance components associated with them will inflate the tolerance intervals associated with predictions in an appropriate way.

The property that random terms have of inducing correlations among the observations is possibly the most important effect of a mixed effects model, as they allow a measure of data integration to take place in the analysis, the so-called ‘borrowing strength’ property. This is particularly important in situations where the data set has not been collected for the primary purpose at hand, but has been drawn together from historical data sets that were originally collected for other purposes. This is the rule rather than the exception in fisheries research in our experience. The data sets are often very unbalanced and even the experimental protocols are changing as time proceeds in perhaps subtle and unknown ways. All these possibilities point to modelling the situation using random terms.

7. Summary

This overview of theory has not attempted to be comprehensive but has instead tried to focus on issues that we find are perennial in fisheries research.

We began with a view of the mathematical genesis of a linear model that we hope made it clear that most linear models are empirical and local in practice, in the sense that they are not expected to perform well outside a restricted domain centred on the observations. With a first-order model, ordinary least-squares normal theory models might be adequate, at least for a continuous response variable. We might expect that interactions among predictors, curvature terms, variance heterogeneity and non-normality will start to play an increasingly important role as the domain of applicability is extended.

If the simple assumptions underlying the normal assumption are not met (well enough), one way of correcting the situation is to transform the response. The goal of a transformation has classically been to stabilise the variance. Even so, a transformation will also change the relationship between the linear predictor and the mean as well as possibly promote overall normality by reducing skewness and kurtosis. Transforming the response may even complicate things if it is necessary to make inference on the mean of the untransformed scale.

Extending the region of applicability of the model may involve including higher-degree polynomial terms in the predictor variables, as suggested by the Taylor series analogue. Other ways of doing this are usually preferable. In particular, regression splines offer a simple way of modelling the dependence of the mean on a predictor variable that offers greater local flexibility. There is still the need, however, either to include cross-product terms or to choose predictor variables in a way that would minimise the need for such terms. This requires an intimate knowledge of the context.

Generalized linear models offer a way of modelling in the original scale, but effectively of accommodating a greater range of (a) links between linear predictor and mean and (b) forms of dependence of the variance on the mean, than is possible under simple normal theory. Inference methods in generalized linear models mainly use the concept of the deviance, which is somewhat like $-2 \times$ log-likelihood, but differs in some essential respects.

Quasi-likelihood models form a non-parametric extension of the idea. They only require a link function, variance function and deviance increment to be supplied and the analysis can proceed even without formal identification of an underlying distribution.

Generalized additive models offer a way of gaining extra flexibility in the linear predictor while automatically controlling the parametric complexity. GAMs allow the linear predictor to include local smoothers as terms. These then entail an estimation procedure that differs from maximum likelihood in that a penalized likelihood is maximized, where the penalty uses a measure of roughness in the fitted surface. The tuning parameter that determines the compromise between fit and roughness is often chosen by cross-validation. These methods offer a powerful exploratory tool in fisheries research, but we have seen few occasions when using appropriately chosen fixed spline terms in the regression is not sufficient. There is also, in practice, a strong incentive to ignore cross-product terms in the linear predictor, because to do so makes the interpretation much simpler and the software facilities provided almost encourages this omission. In our view this avoidance of cross-product terms is a trap. There is a strong requirement for the user of GAMs first to choose variables where the need for cross-product terms is unlikely to be strong (or to include such terms in the model). This is usually not easy and requires some insight into the context.

Generalized linear mixed models are a powerful extension of GLMs. The linear predictor now contains both fixed and random terms. The model is within the GLM family conditional on the random terms, but marginally (i.e. unconditionally), it usually is not. Estimation in GLMMs is still a research topic, and the currently available methods all use some approximation to the multiple integral that defines the marginal distribution in a way that avoids its explicit evaluation.

Mixed effects models offer an important way of using models that are flexible but at the same time have their parameterization strongly controlled. Typically a group of related parameters enters the model as a single random term. The model now focuses not on the separate parameters as would be the case for the fixed effects model but on the variance component, that is, the variance of the distribution from which the parameters are assumed to come. The estimation procedure generally produces ‘predictors’ (often called ‘best linear unbiased predictors’, or BLUPs) of the individual values, although they now have a different logical status from separate parameter estimates. They are, in fact, somewhat analogous to residuals. The BLUPs of the random effects, like the (conditional) residuals, can be subjected to a range of diagnostics such as normal QQplots since the random effects are assumed to be normally distributed in the GLMM. Detection of atypical values of the random effects may also be possible using scatterplots of the BLUPs which include their approximate confidence bounds.

In modelling key parameters as random, mixed effects models have the capacity to ‘borrow strength’ from different parts of the data set and produce predictors of the individual terms that usually show some ‘shrinkage’ towards the general mean, which is seen as natural and reasonable. In other words, the predictors are often much more stable than individual parameter estimates would be, because those use more information in the data. This property makes mixed effect models very effective in situations where the data are very unbalanced or fragmentary, which in turn can result if the data set itself has been put together from historical data sets originally collected for other purposes.

Acknowledgements

The authors wish to thank André Punt and Yongshun Xiao for the invitation to write this paper. We especially appreciate the comments on a draft of the manuscript by Malcolm Haddon, Nick Ellis, Peter Toscas, Mark Bravington and Janet Bishop. We also wish to thank three anonymous referees for many useful suggestions which have led to very useful improvements.

References

Agnew, D.J., Marlow, T.R., Lorenzen, K., Pompert, J., Wakeford, R.C., Tingley, G.A., 2003. Influence of Drake Passage oceanography on the parasitic infection of individual year-classes of southern blue whiting *Micromesistius australis*. Mar. Ecol. Prog. Ser. 254, 281–291.

Bannerot, S.P., Austin, C.B., 1983. Using frequency distributions of catch per unit effort to measure fish stock-abundance. Trans. Am. Fish. Soc. 112, 608–617.

Bigelow, K.A., Boggs, C.H., He, X., 1999. Environmental effects on swordfish and blue shark catch rates in the US North Pacific longline fishery. Fish. Oceanogr. 8, 178–198.

Boneva, L.I., Kendall, D., Stefanov, I., 1970. Spline transformations: three new diagnostic aids for the statistical data analyst. J. Roy. Stat. Soc. Ser. B. 32, 1–71.

Borchers, D.L., Buckland, S.T., Priede, I.G., Ahmadi, S., 1997. Improving the precision of the daily egg production method using generalized additive model. Can. J. Fish. Aquat. Sci. 54, 2727–2742.

Brandão, A., Butterworth, D.S., Johnston, S.J., Glazer, J.P., 2004. Using a GLMM to estimate the somatic growth rate for male South African west coast rock lobster Jasus lalandi. Fish. Res. 70, 335–345.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88, 9–25.

Bromley, P.J., 2000. Growth, sexual maturation and spawning in central North Sea plaice (*Pleuronectes platessa* L.), and the generation of maturity ogives from commercial catch data. J. Sea Res. 44, 27–43.

Brynjarsdóttir, J., Stefánsson, G., 2004. Analysis of cod catch data from Icelandic groundfish surveys using generalized linear models. Fish. Res. 70, 195–208.

Cardinale, M., Arrhenius, F., 2000. The influence of stock structure and environmental conditions on the recruitment process of Baltic cod estimated using a generalized additive model. Can. J. Fish. Aquat. Sci. 57, 2402–2409.

Cooke, J.G., 1997. A procedure for using catch-effort indices in bluefin tuna assessments (revised). ICCAT Col. Vol. Sci. Pap. 46 (2), 228–232.

Cox, D.R., Snell, E.J., 1968. A general definition of residuals (with discussion). J. Roy. Stat. Soc. Ser. B 30, 248–275.

de Boor, C., 1978. A Practical Guide to Splines. Springer-Verlag, New York.

Denis, V., Lejeune, J., Robin, J.P., 2002. Spatio-temporal analysis of commercial trawler data using general additive models: patterns of Loliginid squid abundance in the north-east Atlantic. ICES J. Mar. Sci. 59, 633–648.

Diggle, P., Liang, K.-Y., Zeeger, S.L., 1994. Analysis of Longitudinal Data. Oxford University Press, Oxford.

Finney, D.J., 1941. On the distribution of a variate whose logarithm is normally distributed. Suppl. J. Roy. Stat. Soc. (Ind. Agric. Res. Sec.) VII(2) 104, 155–161.

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

Firth, D., 1987. On the efficiency of quasi-likelihood estimation. Biometrika 74, 233–245.

Firth, D., 1988. Multiplicative errors: lognormal or gamma. J. Roy. Stat. Soc. Ser. B 50, 266–268.

Fisher, R.A., 1954. The analysis of variance with various binomial transformations. Biometrics 10, 130–139.

Fox, J., 1997. Applied Regression, Linear Models, and Related Methods. Sage Publications, Thousand Oaks.

Godambe, V.P., Heyde, C.C., 1987. Quasi-likelihood and optimal estimation. Int. Stat. Rev. 55, 231–244.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.

Kimura, D.K., 1981. Standardized measures of relative abundance based on modelling log(c.p.u.e.), and the application to Pacific ocean perch (*Sebastes alutus*). J. Cons. Int. Explor. Mer. 39, 211–218.

Lai, H-L., Helser, T., 2004. Linear mixed-effects models for weight–length relationships. Fish. Res. 70, 373–383.

Laird, N.M., 1996. Longitudinal panel data: an overview of current methodology. In: Cox, D.R., Hinkley, D.V., Barndorff-Nielsen, O.E. (Eds.), Time Series Models in Economic, Finance and Other Fields. Chapman & Hall, London, pp. 143–175.

Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. 70, 141–149.

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

Myers, R.A., Hoenig, J.M., 1997. Direct estimates of gear selectivity from multiple tagging experiments. Can. J. Fish. Aquat. Sci. 54, 1–9.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. Roy. Stat. Soc. Ser. A 135, 370–384.

Olsen, E., 2002. Errors in age estimates of North Atlantic minke whales when counting growth zones in *Bulla tympanica*. J. Cet. Res. Manage. 4, 185–191.

Ortiz, M., Legault, C.M., Ehrhardt, N.M., 2000. An alternative method for estimating bycatch from the U.S. shrimp trawl fishery in the Gulf of Mexico, 1972–1995. Fish. Bull. US 98, 583–599.

Ortiz, N., Arocha, F., 2004. Alternative error distribution models for standardization of catch rates of non-target species from a pelagic longline fishery: billfish species in the Venezuelan tuna longline fishery. Fish. Res. 70, 275–294.

Punt, A.E., Walker, T.I., Taylor, B.L., Pribac, F., 2000. Standardization of catch and effort data in a spatially-structured shark fishery. Fish. Res. 45, 129–145.

Punt, A.E., Smith, D.C., Thomson, R.B., Haddon, M., He, X., Lyle, J.M., 2001. Stock assessment of the blue grenadier *Macruronus novaezelandiae* resource off south-eastern. Mar. Freshw. Res. 52, 701–717.

Quinn, T.J., 1985. Catch-per unit effort: a statistical model for Pacific halibut. Can. J. Fish. Aquat. Sci. 42, 1423–1429.

R Development Core Team, 2003. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/.

Richards, L.J., Schnute, J.T., 1992. Statistical models for estimating CPUE from catch and effort data. Can. J. Fish. Aquat. Sci. 49, 1315–1327.

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects (with comments). Stat. Sci. 6, 15–32.

Schall, R., 1991. Estimation in generalized linear models with random effects. Biometrika 78, 719–727.

Squires, D., Kirkley, J., 1999. Skipper skill and panel data in fishing industries. Can. J. Fish. Aquat. Sci. 56, 2011–2018.

Stefánsson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. 53, 577–588.

van Dyk, D.A., 2000. The nested EM algorithm. Stat. Sinica 10, 203–225.

Venables, W.N., Dichmont, C.M., 2004. A generalized linear model for catch allocation: an example of Australia's Northern Prawn Fishery. Fish. Res. 70, 405–422.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer-Verlag, New York.

Wahba, G., 1990. Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. Society for Industrial and Applied Mathematics (SIAM), New York.

Walsh, W., Kleiber, P., 2001. Generalized additive model and regression tree analysis of blue shark (*Prionace glauca*) by the Hawaii-based longline fishery. Fish. Res. 53, 115–131.

Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61, 439–477.

Wiens, B.L., 1999. When lognormal and gamma models give divergent results: a case study. Am. Stat. 53, 89–93.

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

Williams, D.A., 1987. Generalized linear model diagnostics using the deviance and single case deletions. Appl. Stat. 36, 181–191.

Wolfinger, R., O'Connell, M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. J. Statist. Comp. Sim. 48, 233–243.

Wright, P.J., Jensen, H., Tuck, I., 2000. The influence of sediment type on the distribution of the lesser sandeel *Ammodytes marinus*. J. Sea Res. 44, 243–256.