Check for updates

European Journal of Soil Science, June 2002, **53,** 241–251

Generalized linear models in soil science

P. W. LANE
Research Statistics Unit, GlaxoSmithKline, Harlow CM19 5AW, UK

Summary

Classical linear models are easy to understand and fit. However, when assumptions are not met, violence should not be used on the data to force them into the linear mould. Transformation of variables may allow successful linear modelling, but it affects several aspects of the model simultaneously. In particular, it can interfere with the scientific interpretation of the model. Generalized linear models are a wider class, and they retain the concept of additive explanatory effects. They provide generalizations of the distributional assumptions of the response variable, while at the same time allowing a transformed scale on which the explanatory effects combine. These models can be fitted reliably with standard software, and the analysis is readily interpreted in an analogous way to that of linear models. Many further generalizations to the generalized linear model have been proposed, extending them to deal with smooth effects, non-linear parameters, and extra components of variation. Though the extra complexity of generalized linear models gives rise to some additional difficulties in analysis, these difficulties are outweighed by the flexibility of the models and ease of interpretation. The generalizations allow the intuitively more appealing approach to analysis of adjusting the model rather than adjusting the data.

Introduction

The linear model is a two-edged sword. Scientists find linear models useful to describe simple relationships between variables, and the models are easy to interpret. In addition, they are straightforward to fit to observations and are implemented in most computer packages that provide for the analysis of data. However, this ease of use results in their application in inappropriate circumstances. There is a tendency to force all modelling problems into the mould of the linear model rather than to make use of more complex models that could actually describe and quantify relationships in a scientifically more meaningful way. For example, soil scientists make use of transformations to achieve approximate normality, with constant variance, of variables that they want to model, in the knowledge that normality and constant variance are two of the underlying assumptions of the linear model. While this adjustment of the data can result in appropriate linear models to represent relationships, it can also lead to difficulties with other aspects of the model, which are much better dealt with by adjusting the form of the model itself.

There is a wider class of models, called generalized linear models, which provide a richer choice. With modern statistical packages, they, too, can be fitted easily, and interpretation is analogous to that of linear models. These models do not appear to have been much used in soil science. However, an early example was provided by Honeysett & Ratkowsky (1989) who modelled the relationship between soil bulk density and low-temperature ignition-loss. They found that an inverse-linear model allowed the retention of a linear relationship while achieving approximate constancy of variance. McKenzie & Austin (1993) used the models to predict soil characteristics such as exchangeable sodium percentage from medium- and small-scale surveys, based on more readily observed environmental variables. More recently, Malafant *et al.* (1999) used a log-linear model, which is one of the best-known types of generalized linear model, to analyse relationships between soil degradation classes and soil physical and chemical limitations.

Transformation of the response variable

When there are difficulties with the assumptions in a classical linear model, a popular option is to recast the model in terms of a transformation of the response. For example, suppose that in simple linear regression it has been found that the variance of the errors increases with the magnitude of the response. This can be signalled by the pattern of residuals plotted against fitted values after a linear analysis, as in the example in the next section. The original model is that the observed responses, y_i, $i = 1, 2, \ldots, n$, are normally distributed with mean and variance given by

$$\mathrm{E}[y_i] = a + bx_i \qquad \text{and} \qquad \mathrm{var}[y_i] = s^2, \tag{1}$$

This paper is one of a series, 'Mathematical Geology – Series for Students', sponsored by the International Association of Mathematical Geology and the *European Journal of Soil Science*.
E-mail: peter.w.lane@gsk.com
Received 15 February 2001; revised version accepted 21 August 2001

© 2002 Blackwell Science Ltd

where x_i is observation i of a single explanatory variable, a and b are intercept and slope parameters, and s^2 is the constant variance. Use of a transformation such as the logarithm may be tried to stabilize the variance – see Webster (2001) in this Journal, though note his warning on confidence limits on page 333. The revised model is that $\log(y_i)$ is normally distributed with mean and variance given by

$$\mathrm{E}[\log(y_i)] = a' + b'x_i \qquad \text{and} \qquad \mathrm{var}[\log(y_i)] = s'^2. \tag{2}$$

It may well be that $\log(y_i)$ does have an approximately normal distribution with constant variance, though this should be checked in the same way as in the untransformed analysis. However, the form of the relationship between Y and X has also been changed, from a straight line to an exponential curve:

$$y_i = a'' \exp(b'x_i)e_i, \tag{3}$$

where $a'' = \exp(a')$ and the errors e_i are multiplicative. There is no difficulty with using the transformed model if a relation of this form is actually suitable; but if a linear relationship is preferred then the transformed model cannot provide a good analysis.

The effect of the transformation may be partly countered by taking logarithms of x_i as well, giving the model

$$y_i = a'' x_i^{b'} e_i. \tag{4}$$

But here the parameter a'' has become the slope of the line, and the parameter b' has to be fixed at 1.0 to make the line straight. There is no intercept, so the line is constrained to pass through the origin. Another parameter can be introduced to give an unconstrained line, but it needs to be inside the log function applied to X. The model

$$\mathrm{E}[\log(y_i)] = a' + \log(x_i + c) \tag{5}$$

is equivalent to

$$y_i = (a''x_i + c')e_i, \tag{6}$$

where $c' = a''c$. The parameter c is not linear within the model, so it cannot be estimated by the methods of the classical linear model; this approach will not be taken any further here.

A similar problem is met if a transformation of the response is used to achieve linearity. If the logarithm of the response is considered to be linearly related to the explanatory variable then the classical model is Equation (2) again, which also requires the distribution of y_i to be log-normal rather than normal with constant variance. The use of a transformation in both these situations suffers from the same drawback: improvements with respect to one assumption are inextricably linked to complications with respect to another.

Another example of the classical linear model is used to analyse the relationship of a response variable with a single factor, or categorical explanatory variable – referred to as a one-way analysis of variance. Suppose that the explanatory variable X has m classes, so that each observation x_i can be represented simply as one of the integers 1, 2, …, m, standing for the number of the class. Then the model can be represented in a form similar to Equation (1):

$$\mathrm{E}[y_i] = b_{x_i} \qquad \text{and} \qquad \mathrm{var}[y_i] = s^2, \tag{7}$$

where b_j, $j = 1, 2, \ldots, m$, is the mean response for class j. If there is a need to transform y_i to stabilize the variance then there is no problem here because the model becomes

$$\mathrm{E}[\log(y_i)] = b'_{x_i} \qquad \text{and} \qquad \mathrm{var}[\log(y_i)] = s^2, \tag{8}$$

and this allows exactly the same pattern of class means as in Equation (7) if $b'_{x_i} = \log(b_{x_i})$.

However, as soon as the one-way model is extended in any way, problems appear again. For example, if a linear contrast of the class comparisons is included, as in the example in the next section, the same problem appears as with linear regression. If a second categorical variable is included, leading to two-way analysis of variance, the problem manifests itself in terms of the additivity of effects. If no interaction is assumed between the two categorical variables, X and Z say, the model can be written as

$$\mathrm{E}[y_i] = b_{x_i} + c_{z_i} \qquad \text{and} \qquad \mathrm{var}[y_i] = s^2. \tag{9}$$

The transformed version is

$$\mathrm{E}[\log(y_i)] = b'_{x_i} + c'_{z_i} \qquad \text{and} \qquad \mathrm{var}[\log(y_i)] = s^2. \tag{10}$$

This imposes a pattern of class means for y_i of the form $b_{x_i}c_{z_i}$, and so is a multiplicative rather than an additive model on the scale of the response. The interpretation of an interaction is completely different in additive and multiplicative models, so the choice of transformation affects the decision of whether or not to include an interaction.

There are many other transformations apart from the logarithm that have been used to try to squeeze data into the mould of the classical linear model. A common example is the square-root transformation, which has the property of approximately stabilizing variance that is proportional to the mean (rather than to the square of the mean for the logarithm). Many measurements that are in the form of counts of objects or events are likely to have this sort of variance–mean relationship. Another transformation is the angular, $\mathrm{arcsine}(\sqrt{y_i/N})$, which approximately stabilizes a variance–mean relationship like that of the binomial distribution. This is relevant for observations that are counts of successes out of a number of trials, but can also be used for continuous measurements that are restricted to a range. If y_i is the number of successes from N_i trials that have probability of success p_i, the binomial variance of y_i is $N_ip_i(1-p_i)$, which is small for probabilities at either end of the range [0, 1].

For these transformations the problems discussed above are compounded with the difficulty of interpretation of additivity on the transformed scale. There is no longer a simple shift from additive to multiplicative effects, as with the logarithm:

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

additive effects on the angular scale cannot be interpreted in any meaningful way in terms of the actual measured observations. However, it remains the case that if all that is needed is an empirical model with few parameters, then these transformations may serve the purpose. They can be used to provide evidence of qualitative effects, as long as the form of the model is not intended to be interpretable scientifically.

The model using the logarithmic transformation is indeed a special case. It has proved in many applications to be useful when at the same time effects are expected to be multiplicative and the variance is expected to be proportional to the square of the mean. But even in such applications the approach of transforming the data to suit the model has practical difficulties which impede the analysis. There has been considerable discussion about what to do with zero measurements when fitting such a model: the logarithm of zero is infinite, so some sort of 'fudge' has to be adopted. In practice, many people use the transformation $\log(y_i + 0.5)$, or $\log(y_i + c/2)$ where c is the smallest non-zero observation. This actually compromises the relationship between the variance and the mean and distorts the multiplicative pattern of effects, particularly for observations that are close to zero. In some applications, the choice of the constant to add before taking logarithms can have critical consequences for the interpretation of an analysis.

A second difficulty is with back-transformation. After analysing a variable on a transformed scale, there is often a need to present results on the original untransformed scale. This requires the back-transformation of results, and it has the unfortunate consequence of producing statistics that do not behave in the same way as from an untransformed analysis. For example, confidence intervals become non-symmetrical, and standard errors may no longer be a good summary of variation. More awkward still, any means that summarize the model for the classes of an explanatory variable will no longer correspond to the means of the raw data, because of the fact that a back-transformed mean of a transformed variable is not the same as the mean of the original variable.

An example of transformation

To illustrate the problems with transformations, I analyse the results of a study on the nitrogen content of soil, carried out at Rothamsted by Glendining *et al.* (1992). I have chosen this example because it involves a small and easily understood set of data and at the same time illustrates many of the features of generalized linear models. The main aim was to describe the relationship between nitrogen applied as fertilizer, at different rates in separate plots over a long period, and the measured nitrogen content from soil samples, taking into account potential differences between inorganic and organic fertilizers. There were three replicates from each of eight plots (Table 1). Six of the plots received inorganic nitrogen fertilizer over a long period, a seventh received organic fertilizer (farmyard manure, FYM), and the last was a control receiving no applications of nitrogen. There is a difficulty in analysing the results of this and many other long-term fertilizer experiments, because the treatments are not replicated. The best that could be done to increase information was to take replicate observations from each plot and assume that the large size of the plots (28 m × 6 m) allows these to be treated as if they were from replicate plots (Glendining *et al.*, 1992). In the analysis that follows I shall ignore this potential difficulty.

Table 1 Soil nitrogen content from eight plots of the Broadbalk experiment

	Replicate		
Treatment	1	2	3
Control	4.53	5.46	4.77
48 kg N ha^{-1}	6.17	9.30	8.29
96 kg N ha^{-1}	11.30	16.58	16.24
144 kg N ha^{-1}	24.61	18.20	30.03
192 kg N ha^{-1}	21.94	29.24	27.43
240 kg N ha^{-1}	46.74	38.87	44.07
288 kg N ha^{-1}	57.74	45.59	39.77
FYM (248 kg N ha^{-1})	25.28	21.79	19.75

The range of the observations here (6.17, 57.74) should be an immediate signal that there is likely to be a problem with the variance. It is difficult to imagine how measurements of this sort that differ by a factor of almost 10 can be subject to the same magnitude of error. However, for the sake of illustration I shall fit a linear model with constant variance to these observations in the first instance. I include the main effect of the nitrogen treatment factor, subdivided into a linear contrast for the amount of nitrogen, the difference between the organic treatment and the rest, and the remainder of the main effect – which represents non-linearity of the nitrogen effect. Table 2 shows the analysis of variance.

This shows that the linear model seems to fit well (the deviations term is non-significant) and that there is a substantial difference between organic and inorganic treatments (about 18 units less soil N than expected from the N content of the manure). However, it is clear even from the information about large residuals that the assumption of constant variance does not hold, because the two largest residuals are both from the treatment with greatest response. A graph of the residuals against fitted values (Figure 1) shows the clear diagnostic pattern of variance increasing with the mean, even though there are not many observations in this example. In fact, the pattern is also apparent in a graph of the original data (Figure 2), so should be picked up before any analysis is done.

A graph of the standard deviation against the mean at each treatment level (Figure 3) shows a roughly linear relationship, so that the variance is roughly proportional to the square of the mean. There are too few data in this example to give much information about the variance–mean relationship, but the graph encapsulates what information there is. I therefore

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2 Analysis of variance of the untransformed response with a linear model

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio	Probability
Treatment	7	4944.5	706.4	33.45	< 0.001
Linear N	1	4000.4	4000.4	189.43	< 0.001
FYM vs N	1	754.0	754.0	35.70	< 0.001
Deviations	5	190.1	38.0	1.80	0.17
Residual	16	337.9	21.1		
Total	23	5282.4	229.7		

Observations with large residuals (greater than twice their standard error):

Treatment	Replicate	Residual (studentized)
288 kg N ha^{-1}	1	2.68
288 kg N ha^{-1}	3	−2.11

Estimates of effects:

Effect	Estimate	Standard error
Intercept (inorganic)	1.59	1.58
Slope (inorganic)	0.15659	0.00912
FYM vs N	−18.15	2.65

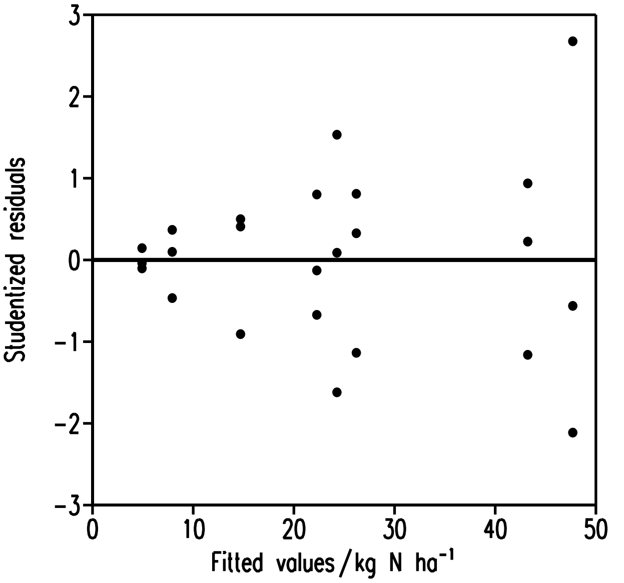

Figure 1 Residuals plotted against fitted values from the untransformed analysis.

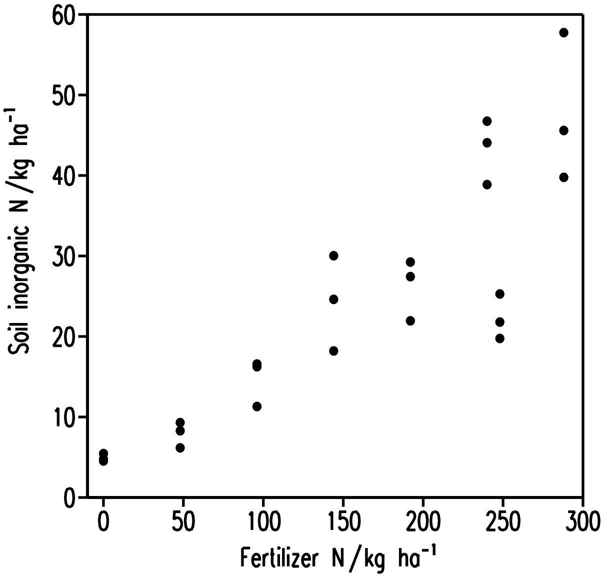

Figure 2 Observed values plotted against amount of fertilizer added.

transform the response using common logarithms to try to stabilize the variance (the natural logarithm could also be used, giving the same analysis but with estimates for the intercept and slope on the transformed scale multiplied by ln(10), i.e. by a factor of about 2.3). The results are summarized in Table 3 in the same way as before.

The problem with non-constant variance seems to be solved, with no large residuals, and lack of relationship between size of residual and fitted value (Figure 4). But the price for this is the introduction of non-linearity of the effect of fertilizer N. The deviations contrast is also now statistically significant, though it accounts for less than 6% of the variance attributable to the treatments. Figure 5 shows the fitted model, which has pronounced curvature on the natural scale. The fitted curve seems to be too low in the middle of the range and too high for large applications of fertilizer N.

Generalized linear models

A solution to all the problems raised above is to modify the model rather than to transform the data. The result of one set of modifications is called the generalized linear model (GLM). It is an attractive candidate for modelling because it shares

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

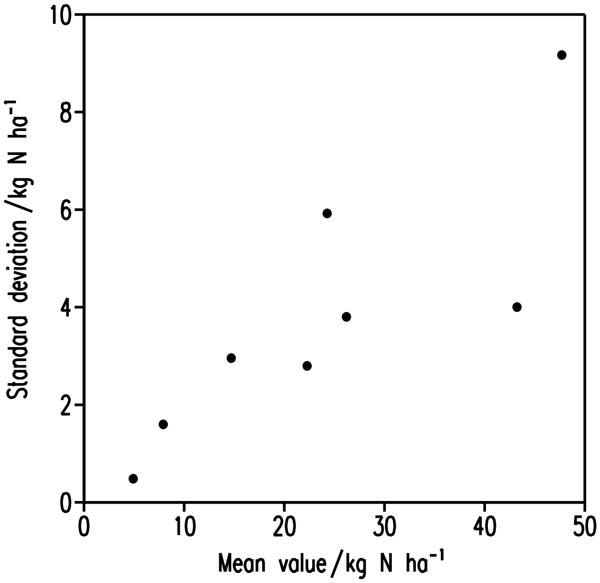

Figure 3 Relationship between standard deviation and mean of the observed data.

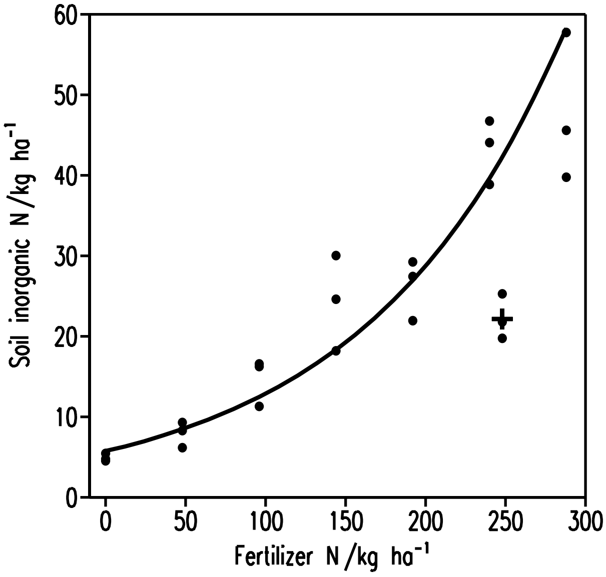

Figure 5 Observations and fitted model using a log transformation; the line shows the fitted effect of inorganic treatment, and + shows the fitted effect of the organic treatment.

many familiar characteristics with the linear model while being appropriate for a much wider range of circumstances. It has become an accepted tool in statistical analysis in many areas of application, but usually only by statisticians. Generalized linear models were first introduced under this name by Nelder & Wedderburn (1972), when the increasing power of computers had made it easy to use more complicated models than linear regression. They are suitable for the same type of problem as linear regression or analysis of variance: modelling

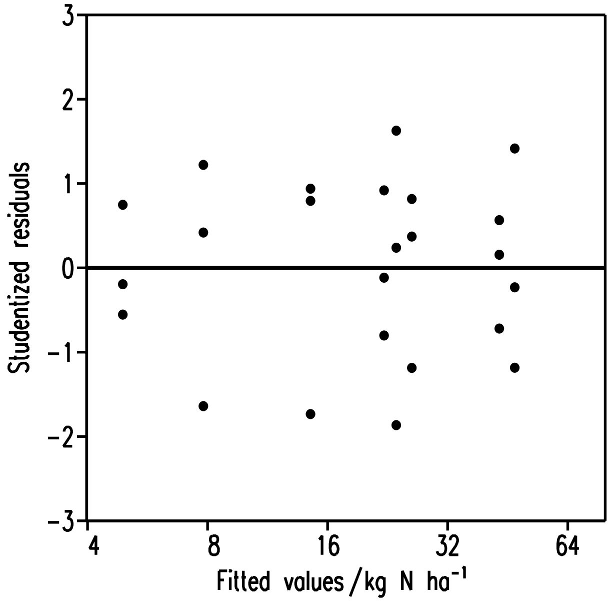

Figure 4 Residuals plotted against fitted values from the log-transformed analysis.

the relationship of one variable (the response or dependent variable) with one or more others (the explanatory or independent variables) to predict the response from the explanatory variables.

In fact, linear regression is the simplest example of a generalized linear model, and other examples had already been developed before 1972. The earliest is probably the development of a method to analyse dilution assays by Fisher (1922). The method of probit analysis for the results of bioassays with quantal responses was developed by Bliss (1935) and Finney (1971); logistic models were also first applied in bioassay by Berkson (1944), and later extended to more general analysis of proportions by Dyke & Patterson (1952). Log-linear models for data in the form of counts were developed in the 1960s, though they were not well known until publication of a book by Bishop *et al.* (1975). Table 4 lists the common forms of GLM. But the main contribution of Nelder & Wedderburn was to show that these examples, previously developed and considered as separate models, were members of a wider class which could be fitted by a common algorithmic technique and applied with a common strategy. Since 1972, the methodology has spread to a very wide range of applications, and has slowly become available in most of the major statistical packages. Table 5 lists the packages that provide a general facility for fitting GLMs, though many more provide individual models separately. The comprehensive book by McCullagh & Nelder (1989), and a simple introduction by Dobson (1990), have also contributed to the popularity of the techniques.

Despite the acceptance of GLMs by practitioners, there are many scientists who have either not heard of them or have not taken the trouble to find out what they offer. Indeed, there is

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3 Analysis of variance of the $\log_{10}$-transformed response with a linear model

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio	Probability
Treatment	7	2.47071	0.35296	60.66	< 0.001
Linear N	1	2.19067	2.19067	376.49	< 0.001
FYM vs N	1	0.18001	0.18001	30.94	< 0.001
Deviations	5	0.10003	0.02001	3.44	0.027
Residual	16	0.09310	0.00582		
Total	23	2.56381	0.11147		

Observations with large residuals (greater than twice their standard error):
None

Estimates of effects:

Effect	Estimate	Standard error
Intercept (inorganic)	0.7612	0.0262
Slope (inorganic)	0.003487	0.000151
FYM vs N	−0.2804	0.0440

Table 4 Common generalized linear models

Common name	Response variable	Explanatory variables	Link function	Error distribution
Linear regression	Continuous	Continuous	Identity	Normal
ANOVA	Continuous	Categorical	Identity	Normal
ANCOVA	Continuous	General	Identity	Normal
Inverse polynomials or inverse-linear regression	Continuous	Continuous	Inverse	Normal
Log-linear model or Poisson regression	Counts	Categorical	Log	Poisson
Probit analysis	Proportions	Continuous	Probit	Binomial
Logistic regression	Binary	General	Logit	Binomial
Dilution assay	Proportions	Continuous	Complementary log–log	Binomial
Survival analysis	Continuous	General	Inverse	Exponential

Table 5 Software to fit GLMs

Package	Section	Internet address
Genstat	All regression commands	vsn-intl.com/genstat/
GLIM	All model-fitting commands	www.nag.co.uk/stats/GDGE_soft.asp
SAS	Genmod procedure	www.sas.com/products/
S-plus	GLM function	www.insightful.com/products/splus/
Stata	GLM command	www.stata.com/
Statistica	GLZ module	www.statsoft.com/

still confusion about the abbreviation GLM, generally adopted by users of generalized linear models, but still in use by others for the much more restrictive general linear model, which includes only linear techniques such as linear regression and analysis of covariance. This does not help people new to the methodology, since they may take references to GLMs to mean no more than linear models, and look no further.

The need for GLMs is apparent when the main assumptions of classical linear models cannot be made. Specifically, these are that the explanatory variables combine additively in their effects on the response, the response variable is distributed normally (at least approximately), and the variance of the responses is constant and hence independent of the mean. Other assumptions in classical linear models, such as that the responses

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

are distributed independently, are also needed for standard GLMs.

The GLM provides two separate modifications to the classical linear model which address the two main problems of constant variance and additivity. The first is to allow the distribution of the responses to be any of the wide exponential family of distributions. This family includes the normal, where the variance is independent of the mean, and also the Poisson (variance equal to mean), gamma (variance proportional to the square of the mean), and the binomial (variance equal to $Np(1-p)$, where the mean is Np). In two of these distributions, the normal and the gamma, there is an additional dispersion parameter, ϕ say, which must be assumed constant over all observations, or have a known set of relative weights for each observation, w_i say. This is a familiar concept in classical linear models, where ϕ is the variance of the normal distribution, usually denoted by σ^2, and observations may be weighted. With the gamma distribution, the dispersion parameter is usually referred to as the square of the coefficient of variation, also denoted by σ^2; the reciprocal of the dispersion parameter is the index of the gamma distribution.

With this modification alone, the linear regression model becomes

$$\mathrm{E}[y_i] = a + bx_i \qquad \text{and} \qquad \mathrm{var}[y_i] = v(\mathrm{E}[y_i]), \qquad (11)$$

where the function $v()$ is the variance–mean relationship of a distribution in the exponential family. For example, with the gamma distribution:

$$\mathrm{var}[y_i] = \sigma^2\mathrm{E}^2[y_i]. \qquad (12)$$

The second modification is to introduce a link function which relates the mean of the response to a scale on which the effects in the model combine additively. The identity function (equivalent to omitting a function altogether) corresponds to the assumptions of the classical linear model, but any other monotonic function can be used, including the logarithm, square root, logistic and even the angular transformation. The linear regression model now becomes

$$f(\mathrm{E}[y_i]) = a + bx_i \qquad \text{and} \qquad \mathrm{var}[y_i] = v(\mathrm{E}[y_i]), \qquad (13)$$

where $f()$ is the link function. This is often expressed using the inverse of the link function, $g() = f^{-1}()$, which is the exponent in the case of the log link, for example:

$$\mathrm{E}[y_i] = g(a + bx_i) \qquad \text{and} \qquad \mathrm{var}[y_i] = v(\mathrm{E}[y_i]). \qquad (14)$$

The combination of explanatory effects, here just $a + bx_i$, is referred to as the linear predictor. This can be extended in the same way as in multiple linear regression or analysis-of-variance models to include further effects of continuous or categorical variables and interactions between them. This is one of the key features of GLM methodology: you can combine explanatory effects in the same familiar way as in linear models, while making use of the distribution to deal with properties of the response variable, and of the link function to choose a suitable scale. In matrix notation we can represent the GLM as

$$\mathrm{E}[y] = g(\mathbf{bX}) \qquad \text{and} \qquad \mathrm{var}[\mathbf{y}] = v(\mathrm{E}[\mathbf{y}]), \qquad (15)$$

where the matrix $\mathbf{X}$ has n rows (number of observations) and m columns (number of explanatory effects) and is called the design matrix, as in classical linear models. The parameters are now represented by the vector quantity $\mathbf{b}$.

One consequence of making these generalizations is that the models cannot be fitted using the same calculations as for the linear model. As with any type of statistical modelling there are different approaches to the fitting process, including Bayesian and frequentist methods, but the most common is the likelihood approach which seeks to maximize the likelihood of the observations in terms of the parameters of the chosen model. With the classical linear model, maximum-likelihood corresponds to least-squares, and the parameter estimates can be calculated analytically. But with models that do not assume normal distributions, or that include a non-linear link function, the maximum-likelihood estimators in general cannot be derived analytically, and an iterative computational technique is needed. This is probably the main reason why such models were not popular before modern computing power allowed us to cope with them.

Nelder & Wedderburn (1972) showed that all GLMs could be fitted using a version of the Gauss–Newton iterative algorithm, usually called the scoring algorithm. It converges quickly and reliably in nearly all practical applications, particularly for specific canonical pairs of link function and response distribution that have certain theoretical properties. For example, the canonical link for the Poisson distribution is the log link, and for the binomial it is the logit. Problems can arise when the model is a poor fit to the data and unusual combinations of distribution and link function are chosen. In modern statistical packages, the iterative nature of the underlying fitting process is dealt with automatically without the need for any special action by the user. The scoring algorithm is effectively a method of iteratively re-weighted least-squares. An adjusted response variable is constructed, on the scale of the linear predictor, and a linear model fitted to it using a set of weights. The results of the fit are then used to recalculate the weights from the variance function, and the process is repeated until convergence.

A second consequence of generalizing the linear model is that the familiar summary of a linear model in terms of an analysis of variance cannot be used. Even when the normal distribution is retained, non-linearity in the link function means that the distributional theory for variance ratios becomes approximate (Seber & Wild, 1989, p. 32). With other distributions, variance ratios are not appropriate anyway for expressing the contribution of effects to the fit of a model. However, Nelder & Wedderburn showed that an analogous analysis of deviance can be constructed for any distribution in the exponential family. This, too, is approximate, even if the

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

link function is linear; but in practice, the asymptotic behaviour of deviance statistics is approached rapidly for even modest numbers of observations in many models.

The deviance is a measure of the discrepancy of the observations from the fitted model. It is easiest to define in terms of the scaled deviance, which is minus twice the log-likelihood ratio of the fitted model with a saturated model – a model with the same number of parameters as of observations, so that the fitted values equal the observations. This statistic is otherwise known as the likelihood-ratio statistic. Multiplying the scaled deviance by the dispersion parameter gives the deviance itself, which is thus a quantity that does not depend on the (usually unknown) dispersion parameter. If the distribution is normal, then the deviance is the residual sum of squares; so the analysis of deviance is, reassuringly, the analysis of variance. But for the gamma distribution, for example, the deviance is

$$2\sum_{i=1}^{n}\left\{\frac{y_i - f_i}{f_i} - \log\left(\frac{y_i}{f_i}\right)\right\}, \tag{16}$$

where f_i represents the fitted value corresponding to the observation y_i, and there are n observations. The deviance is thus very different from the residual sum of squares:

$$\sum_{i=1}^{n}(y_i - f_i)^2. \tag{17}$$

However, just like the residual sum of squares it can be divided into components associated with individual terms in the linear predictor, and so the style of sequential analysis familiar in linear models is available. Scaled deviances have approximate χ^2 distributions (exact for linear models with the normal distribution), and so do components of the scaled deviance. When the dispersion parameter is unknown, it can be estimated from the residual deviance: the ratio of the mean deviance of a term to the residual mean deviance then has an approximate F distribution (again, exact for linear models with the normal distribution).

Example of a generalized linear model

In the example of soil nitrogen introduced above, the GLM framework allows us to cater for the variance–mean relationship without detriment to the required linear relationship between soil and fertilizer N. Table 6 shows the results of the analysis.

The deviations term is statistically significant as in the transformed analysis, though again it accounts for only a small proportion, now 7%, of the deviance attributable to the treatment term. The residuals from a GLM can be standardized in a way analogous to those from linear models, allowing the same types of checking of the model. Here, there are no unusually large residuals because the apparently large deviations from the model for large values of the explanatory (reported in the analysis of the linear model) are not unusual under the assumption of a gamma distribution where the variability increases with the mean.

The parameter estimates are similar to those from the linear model. The intercept is larger by about three units and has a much smaller standard error. The estimate of the organic effect is about four units smaller in magnitude, with a similar value of the standard error. The linear effect of fertilizer N is about 17% smaller, and is more precisely estimated (the standard error is about 30% smaller). All these differences can be attributed to the different weighting given to each observation in this analysis as a result of the assumed variance–mean relationship. The differences can easily be much larger in other circumstances: in particular, estimates of variability such as standard errors can be very different in a linear model from those in a model with more appropriate assumptions about the variance function.

Table 6 Analysis of deviance, with a gamma distribution and a linear model

Source of variation	Degrees of freedom	Deviance	Mean deviance	Deviance ratio	Approx. probability
Treatment	7	11.10578	1.58654	60.66	< 0.001
Linear N	1	9.99550	9.99550	330.47	< 0.001
FYM vs N	1	0.56058	0.56058	18.53	< 0.001
Deviations	5	0.54970	0.10994	3.63	0.022
Residual	16	0.48394	0.03025		
Total	23	11.58972	0.50390		

Observations with large residuals (greater than twice their standard error):
None

Estimates of effects:

Effect	Estimate	Standard error
Intercept (inorganic)	4.433	0.370
Slope (inorganic)	0.13037	0.00640
FYM vs N	−14.49	2.45

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The behaviour of the F statistics in a model like this can be checked by simulation. This is a computer-intensive process, and so not suitable for use every time a model is fitted, but it is instructive to see how close is the asymptotic approximation in this illustrative example. I checked the distribution of the mean deviance ratio for the deviations term in the analysis, by analysing 1000 simulations of the data. The response values were generated from gamma distributions with means equal to the values fitted to the original data by a model excluding the deviations term, and with dispersion parameter equal to that observed in the original data with this model. The mean and standard deviation of the simulated ratios were 1.109 and 0.842, whereas the theoretical values for the asymptotic F distribution with 5 and 16 degrees of freedom are 1.143 and 0.910. Thus the ratio appears to be slightly negatively biased by about 3%; but even 1000 simulations are not enough to be confident that this is not just due to chance, because the 95% confidence interval for the mean of 1000 $F_{5,16}$-distributed values is (1.087, 1.199). The asymptotic distribution thus seems to be adequate for practical purposes, even though the number of observations in this example is small. Of course, the situation may be worse in GLMs that also have a non-linear link function as well as a non-normal distribution, but there does not yet seem to be general guidance available on this point.

Extensions of generalized linear models

Since the formulation of GLMs in 1974, many extensions to the original framework have been developed. I give here a brief summary and reference to each of the main areas.

Additive models, previously a separate extension to linear models, were extended to generalized additive models (GAMs) by Hastie & Tibshirani (1990). The extension is to the form of the linear predictor, which is allowed to include smoothed effects of quantitative explanatory variables. Various types of smoothing have been used, of which the most popular are locally weighted regression (referred to as LOESS) and cubic smoothing splines. The classical method of polynomial regression can also be seen as a type of smoothing, but it has a simple parametric form that can readily be interpreted in terms of a linear component, a quadratic component, and so on. The popular smoothing methods have a complex parametric form, and they are not intended to provide an interpretable form for the effect, so the models are often referred to as non-para metric models. They are used to describe an effect graphically, perhaps to suggest a suitable parametric form; or they can just take account of the effect in a general way, suggested by the data rather than a pre-defined form of relationship, in order to concentrate on other more important effects. For example, such a model could be fitted to the example above to take account of the effect of fertilizer N without imposing a linear relationship; the effect of the organic fertilizer could then be estimated with the effect of inorganic fertilizer already accounted for.

Another extension to the framework deals with response variables that have ordinal classes (McCullagh, 1980). For example, soils may be divided into standard classes that can be ordered according to freedom of drainage, with labels such as ‘excessive’, ‘free’, ‘impeded’, and so on. The relationship of this variable to a set of explanatory variables may then need to be modelled. Rather than assigning arbitrary values to each class and analysing the resulting quantitative variable as a continuous response, ordinal response models allow the estimation of cut-points on an underlying continuous scale. This allows differences between classes to be estimated from the data rather than fixed in advance.

Generalized non-linear models (GNMs) allow for extra non-linear parameters within a model that otherwise has a generalized linear form (Lane, 1996). The non-linear parameters may be part of the ‘linear predictor’, which then only has to be linear given the values of the non-linear parameters; this allows non-linear effects of some explanatory variables to be introduced into a model. Alternatively, the link function may involve parameters that have to be estimated from the data. Finding the maximum-likelihood estimates for parameters of such a model requires a non-linear search algorithm. But the separation of the parameters into a (preferably small) set of non-linear parameters and a (relatively large) set of linear parameters within the linear predictor allows a much more efficient and reliable search than would be possible if all parameters had to be treated as non-linear.

When a simple variance–mean relationship is not sufficient to explain the pattern of variation in a set of data, it may be possible to use fixed weights, as in linear models. But without prior knowledge of such weights, the dispersion of the response may be modelled in terms of the explanatory variables at the same time as modelling the location of the response in terms of the explanatory variables. Nelder (1992) showed how such joint modelling of mean and dispersion can be carried out, introducing the idea of extended quasi-likelihood.

A serious impediment to the use of GLMs for analysing data from large planned studies has been the restriction to a single random term in the model. This has meant that count and proportion data from blocked designs, for example, cannot be analysed in a way analogous to the analysis of variance with multiple error terms. Re-expressing Equation (15) to show the actual value of the response rather than its mean gives

$$y = g(\mathbf{bX}) + \mathbf{e}, \tag{18}$$

where **e** is a vector of errors, or residuals. Whereas any number of fixed terms can be included in the linear predictor **bX**, there is only one random term, namely **e**. Various approaches have been suggested for introducing extra random terms. Schall (1991) introduced generalized linear mixed models (GLMMs) in which extra error terms (apart from the term **e**) are included inside the function $g()$, and assumed to have normal distributions. Different algorithms to fit these models have been developed in different areas of application, including social

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

statistics, where the models are referred to as multilevel models (Goldstein, 1995). More recently, Lee & Nelder (1996) developed a more general class of hierarchical generalized linear models (HGLMs) which include GLMMs, but also allow other distributions for the extra error terms.

Advantages and disadvantages of GLMs

The main drawback of using GLMs is that they are unfamiliar and more difficult to explain than linear models. Whereas most people carrying out or using the results of data analysis are familiar with the normal distribution and linear models, many may have little knowledge of distributions such as the binomial and Poisson, far less the gamma, and may have little experience of dealing with non-linearity. This, of course, is no excuse for falling back on a model that demonstrably does not fit well or violates the underlying assumptions of an analysis. Fortunately, the techniques of logistic regression, probit analysis and log-linear modelling have spread into many areas of application. This allows other GLMs, such as the *ad hoc* model developed in the example above, to be explained and justified by reference to generally accepted methodology.

Another cause for concern is the approximation inherent in the use of results based on asymptotic distributions in the analysis of deviance. It is tempting to think that because the distributional results for linear models are exact, they are therefore preferable to the approximations in a more complex model. But this view ignores the fact that the results for linear models are exact only if the underlying assumptions are valid. If there are violations, then statistics such as standard errors and variance ratios are likely to have properties very different from those that are expected in a well-behaved problem, and they are likely to be far more misleading than discrepancies in a more complex model caused by asymptotic approximation. Again, established techniques such as logistic regression show that the approximations are frequently used. Simulation results such as those above show that the approximations are unlikely to be serious, even with a fairly small set of data.

A serious practical difficulty can arise in GLMs with distributions such as the binomial and Poisson in which there is no dispersion parameter. What can be done when all possible effects have been included in the model, but the residual mean deviance still shows a significant χ^2 value? This is evidence either that there are effects that have not been taken into account, or that the variance–mean relationship is not as assumed. Finney (1971) discusses this problem in the context of probit analysis, and adopts a pragmatic solution involving the estimation of a heterogeneity parameter. This takes the place of the dispersion parameter, as with the normal and gamma distribution, but there is actually no theoretical distribution that corresponds to the modified variance–mean relationship. Wedderburn (1974) gave a theoretical basis to this when he developed the concept of quasi-likelihood to deal with the problem. He showed that the variance–mean relationship is the only aspect of a distribution that affects the analysis, and that inference can be based on an assumption just of the form of this relationship with no further specification of the distribution.

In any non-linear problem, maximum-likelihood estimates are biased. For large samples, the bias is negligible compared with standard errors, but for small samples, or models in which the number of parameters is appreciable compared with the number of observations, the bias may be more important (McCullagh & Nelder, 1989, p. 455). For example, to remove bias in logistic regression with a binary response, all parameter estimates need to be shrunk towards the origin by a factor of approximately $(1 - m/n)$, where m and n are the numbers of parameters and observations, respectively. The practical importance of this bias depends on the size of standard errors, but the bias will be less than 5% as long as there are at least 20 observations for each parameter in this model.

A final inconvenience occurs in nearly all GLMs except the classical linear model: explanatory effects are not orthogonal on the scale of the linear predictor even if data collection has been designed to give balance. This means that there is in general no unique deviance attributable to any effect, but each deviance depends on which other effects have already been accounted for and which have not, in the sequence of adding effects to the model. This, of course, is a familiar state of affairs in regression analysis, which is often used for studies where data are acquired in a way that does not allow balance to be achieved. But it is inconvenient to have to use the techniques of regression, looking at alternative sequences of explanatory effects to be sure of inference, in situations that have been carefully designed to give balance.

To outweigh this catalogue of disadvantages, I hope that the example above has illustrated that there are clear advantages of using GLMs, as follows.

- From an intuitive viewpoint, it is better to adjust the model for the data that have been observed rather than to adjust the data to suit a pre-defined model.
- It is unarguable that the class of GLMs offers flexibility, providing scope to deal with a wide range of patterns for the relationship between variance and mean.
- Most crucially, the two generalizations provide the ability to separate the choice of scale on which effects are to be linear and additive from the need to model the variance behaviour of the response; moreover, the option to transform the response and any explanatory variable is also available before fitting a GLM.
- With GLMs there is no problem with interpreting means on the natural scale, because the model is defined in terms of expectations of the response on that scale.
- There is much less difficulty with extreme observations for constrained data, such as zero counts; unless all observations used to estimate a parameter are extreme, no special action is needed when using the Poisson or binomial distribution (though it is not generally possible to deal with zero observations when using the gamma distribution).

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Acknowledgements

The initial work for this paper was done while I was in the Statistics Department, IACR–Rothamsted, and I thank Dr M.J. Glendining for permission to use the data in the illustrative example. I thank Professor R. Webster for encouragement to prepare the paper and advice in presenting it, and the referees for several insightful remarks and suggestions.

References

Berkson, J. 1944. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, **39,** 357–365.

Bishop, Y.V.V., Fienberg, S.E. & Holland, P.W. 1975. *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA.

Bliss, C.I. 1935. The calculation of the dosage–mortality curve. *Annals of Applied Biology*, **22,** 134–167.

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

Dyke, G.V. & Patterson, H.D. 1952. Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8,** 1–12.

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

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Proceedings of the Royal Society*, **222,** 309–368.

Glendining, M.J., Poulton, P.R. & Powlson, D.S. 1992. The relationship between inorganic N in soil and the rate of fertilizer N applied on the Broadbalk Wheat Experiment. *Aspects of Applied Biology*, **30,** 95–102.

Goldstein, H. 1995. *Multilevel Statistical Models*, 2nd edn. Arnold, London.

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

Honeysett, J.L. & Ratkowsky, D.A. 1989. The use of ignition loss to estimate bulk density of forest soils. *Journal of Soil Science*, **40,** 299–308.

Lane, P.W. 1996. Generalized nonlinear models. In: *Compstat Proceedings in Computational Statistics* (ed. A. Prat), pp. 331–336. Physica-Verlag, Heidelberg.

Lee, Y. & Nelder, J.A. 1996. Hierarchical generalized linear models. *Journal of the Royal Statistical Society B*, **58,** 619–656.

Malafant, K.W.J., Atyeo, C.M. & Derbyshire, P.K. 1999. Degradation propensity in Australian land tenure systems. *Land Degradation and Development*, **10,** 455–466.

McCullagh, P. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society B*, **42,** 109–142.

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

McKenzie, N.J. & Austin, M.P. 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. *Geoderma*, **57,** 329–355.

Nelder, J.A. 1992. Joint modelling of mean and dispersion. In: *Sixth International Workshop on Statistical Modelling* (eds P.G.M. van der Heijden, W. Jansen, B. Francis & G.U.H. Seeber), pp. 263–272. Elsevier, Amsterdam.

Nelder, J.A. & Wedderburn, R.W.M. 1972. Generalized linear models. *Journal of the Royal Statistical Society A*, **135,** 370–384.

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

Seber, G.A.F. & Wild, C.J. 1989. *Nonlinear Regression*. John Wiley & Sons, New York.

Webster, R. 2001. Statistics to support soil research and their presentation. *European Journal of Soil Science*, **52,** 331–340.

Wedderburn, R.W.M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61,** 439–447.

13652389, 2002, 2, Downloaded from https://bsssjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2389.2002.00440.x by Georg-August-Universitaet, Wiley Online Library on [18/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License