Biometrika (1991), 78, 1, *pp.* 45-51
Printed in Great Britain

Nonlinear multilevel models, with an application to discrete response data

BY HARVEY GOLDSTEIN
Department of Mathematics, Statistics and Computing, Institute of Education, University of London, London WC1H 0AL, U.K.

SUMMARY

A procedure is proposed for the analysis of multilevel nonlinear models using a linearization. The case of log linear models for discrete response data is studied in detail.

Some key words: Iterative generalized least squares; Log linear model; Multilevel model; Nonlinear model.

1. INTRODUCTION

Nonlinear models arise in a number of circumstances, notably when modelling discrete data. In this paper we consider the multilevel nonlinear model. As in linear multilevel models, we shall consider the general case where any of the model coefficients can be random at any level, and where the random parameters may also be specified functions of the fixed parameter estimates, discussed by H. Goldstein, R. Prosser and J. Rasbash in an as yet unpublished report. In the next two sections we set out the model and define notation; this is followed by a section on estimation and then some examples.

2. THE LINEAR MULTILEVEL MODEL

The general multilevel linear model can be written in the form (Goldstein, 1989)

$$y = X\beta + Ze, \quad E(e) = 0, \quad E\{(Ze)(Ze)^{\mathrm{T}}\} = V,$$

where β is a vector of fixed coefficients, to be estimated, X is the design matrix for the fixed coefficients, e is a vector of random variables and Z is the design matrix for these. The columns of X and Z may or may not coincide.

The matrix V is typically unknown, but for a standard $(h+1)$-level model $(h>0)$ is structured as follows (Goldstein, 1986)

$$V_{h+1} = V_h + Z_{h+1}(I_{(n_{h+1})} \otimes \Omega_{h+1}) Z_{h+1}^{\mathrm{T}},$$

where Ω_{h+1} is the covariance matrix for the model parameters which vary randomly over the $(h+1)$-level units, that is at level $(h+1)$. This equation expresses the matrix for a $(h+1)$ level model recursively in terms of matrices for lower level models. In particular, V_1 is the contribution to V of the level 1 variation and is proportional to the identity matrix in the simplest case. The matrix Z_{h+1} contains the values of the explanatory variables with coefficients random at level $(h+1)$.

3. THE NONLINEAR MULTILEVEL MODEL

The general model can be written as the sum of a nonlinear component and a linear component, both of which may contain fixed and random variables,

$$y = f(X_1\beta + Z_u u) + X_2\gamma + Z_e e, \tag{1}$$

where f is a nonlinear function, e, u are sets of random variables with zero means and corresponding design matrices Z_e, Z_u, β, γ are vectors of fixed coefficients with design matrices X_1, X_2. We write

$$\beta = (\beta_0, \ldots, \beta_l)^{\mathrm{T}}, \quad \gamma = (\gamma_0, \ldots, \gamma_m)^{\mathrm{T}}, \quad u = (u_1, \ldots, u_p)^{\mathrm{T}}, \quad e = (e_1, \ldots, e_q)^{\mathrm{T}},$$

$$X_1 = (x_{11}, \ldots, x_{1l}), \quad X_2 = (x_{21}, \ldots, x_{2m}), \quad Z_u = (z_{u1}, \ldots, z_{up}), \quad Z_e = (z_{e1}, \ldots, z_{eq}).$$

Usually Z_u, Z_e are subsets of X_1, X_2, but they need not be so, and they may contain vectors in common. The random variables may be random at any levels. In many important applications e will contain vectors random at level 1 and u will contain vectors random at higher levels. An example of such a model is that proposed by Jenss & Bayley (1937) for growth in height of children up to the age of six years. For age t this can be written

$$y_{ij} = -\exp(\beta_{0ij} + \beta_{1j}t) + \gamma_{0ij} + \gamma_{1j}t, \quad \beta_{0ij} = \beta_0 + u_{\beta 0j} + e_{\beta ij},$$

$$\beta_{1j} = \beta_1 + u_{\beta 1}, \quad \gamma_{0ij} = \gamma_0 + u_{\gamma 0j} + e_{\gamma ij}, \quad \gamma_{1j} = \gamma_1 + u_{\gamma 1},$$

where i indexes the measurement occasion for the jth individual. Here the design matrices for the random and fixed parts of the models coincide.

In the present paper we consider in detail the particular case of the 2-level log linear model for discrete response data. The extension to further levels is straightforward. The response vector in this model is a vector of proportions, one for each cell (i) of a multiway classification within each level 2 unit. We write

$$\log(\pi_{hij}) = \sum_{k=0}^{l} \beta_{jk} x_{hijk},$$

$$\pi_{hij} = \exp\left(\sum_{k=0}^{l} \beta_{jk} x_{hijk}\right), \quad \sum_i \pi_{hij} = 1 \quad (h = 1, \ldots, q;\ i = 1, \ldots, m_j) \tag{2}$$

for the mean proportion for the hth level 1 unit within the ith cell of the jth level 2 unit. There are a total of qm_j level 1 units in the jth level 2 unit.

Equation (2) describes the mean responses for a generalized linear model with observed responses p_{hij}, where n_{ij} is the size of the ith cell of the jth level 2 unit and typically $n_{ij}p_{hij}$ conditionally have a multinomial distribution with mean $n_{ij}\pi_{hij}$.

In the 2-level model some or all of the coefficients β_{jk} are assumed to be random variables at level 2. Thus

$$\beta_{jk} = \beta_k + u_{jk} \quad (k = 1, \ldots, l). \tag{3}$$

The u_{jk} are continuously distributed level 2 random variables with zero mean and finite covariance matrix. The x_{hijk} consist typically of dummy variables at level 1 defining categories, and covariates at level 2 which are measured at that level; see Goldstein (1987, Ch. 6) for details.

In the following development we first consider the general model (1) and then study specific applications to discrete response data.

4. Estimation

We approach the estimation by considering first a linearization of f, followed by application of a standard procedure for the linear multilevel model. The IGLS algorithm provides a particularly flexible procedure, although the Fisher scoring algorithm of Longford (1987) or the EM procedure of Raudenbush & Bryk (1986) can also be adapted for special cases. All three algorithms give maximum likelihood or restricted maximum likelihood estimates in the normal case (Goldstein, 1986, 1989). Del Pino (1989) also defines an IGLS algorithm, but in his formulation V is a function of the parameters β and he discusses in detail only the case of independent residuals, that is the single level model.

At a given iteration, say $(t+1)$, we assume that values for the fixed coefficients and the random parameters, that is the variances and covariances, are available from the previous iteration. We consider the linear first-order terms of the Taylor expansion for each element of f. Ignoring subscripts we have

$$f = f(H_t) + \sum_{k=1}^{p} h_k (\partial f / \partial h_k)_t. \tag{4}$$

Here H represents the fixed part of the model, $X_1\beta$, and h_k denotes the kth term in $Z_u u$, that is $z_{uk}u_k$, with mean zero.

Consider first the random part of the model, that is the second term in (4) with the expansion about the value $H = H_t$. This can be written as

$$\sum_{k=1}^{p} u_k z^*_{uk}, \quad z^*_{uk} = z_{uk}(\partial f / \partial h_k)_t = z_{uk} f'(H_t). \tag{5}$$

Likewise, the fixed part of the model, the first term in (4), can be written as the Taylor series

$$f(H_t) + \sum_{k=0}^{l} (\beta_{k,t+1} - \beta_{k,t}) x_{1k} f'(H_t) = \left(f_1(X_1\beta) - \sum_{k=0}^{l} \beta_{k,t} x^*_{1k} \right) + \sum_{k=0}^{l} \beta_{k,t+1} x^*_{1k},$$

$$x^*_{1k} = x_{1k} f'(H_t).$$

For the full model we now have

$$y - \left(f_1(X_1\beta) - \sum_{k=0}^{l} \beta_{k,t} x^*_{1k} \right) = \sum_{k=0}^{l} \beta_{k,t+1} x^*_{1k} + \sum_{k=1}^{p} u_k z^*_{uk} + \sum_{k=0}^{m} \gamma_k x_{2k} + \sum_{k=1}^{q} e_k z_{ek}. \tag{6}$$

We now have a standard form of the multilevel linear model and estimates for the fixed and random parameters can be obtained in the usual way. The term in large parentheses on the left-hand side of (6) is updated at each iteration.

Expression (6) has the general form

$$y - \mu_t = X^*_1(\beta_{t+1} - \beta_t) + X_2\Gamma + Z_u u + Z_e e$$

and the adjustment to β_t given by $(\beta_{t+1} - \beta_t)$ is obtained as for the general quasilikelihood model (McCullagh & Nelder, 1983, § 8.5). In the present case the variance matrix is not generally diagonal and needs to be estimated from the data. This is done iteratively, analogously to the ordinary multivariate normal case described above, and the weight matrix used in the estimation of the variances and covariances is that which assumes multivariate normality. While the estimates produced are consistent, their detailed properties are unknown. Nelder & Pregibon (1987, § 3.3) outline a similar approach to variance estimation for the single level case.

The above assumes that the series expansion expression for the random component of (6) is exact rather than an approximation and we return to this issue following some examples.

5. EXAMPLES

The examples are concerned with the analysis of unemployment data in Scotland. The response variable is the proportion of individuals employed and is categorized by gender and qualification level, i.e. unqualified, qualified. A total of 122 geographical areas were sampled and so we have a 2-level model with 4 (2×2) level 1 units per level 2 unit but often with some level 1 units missing. In this case the response vector in (2) has only two categories and the constraint that they sum to one yields a single response variable. The number, m_j, of cells per level 2 unit has a maximum value of 4.

We write a main effects model

$$\pi_{ij} = \exp(\beta_{j0}+\beta_1 x_{ij1}+\beta_2 x_{ij2})\{1+\exp(\beta_{j0}+\beta_1 x_{ij1}+\beta_2 x_{ij2})\}^{-1} \quad (i=1,\ldots,4), \qquad (7)$$

where x_{ij1} is a dummy variable for gender, x_{ij2} is a dummy variable for qualification level and π_{ij} is the expected proportion in the ith cell of the 2×2 classification for the jth area. The intercept term is random, $\beta_{j0}=\beta_0+u_j$. Differentiating, we have

$$\pi'_{ij} = \pi_{ij}\{1+\exp(\beta_{j0}+\beta_1 x_{ij1}+\beta_2 x_{ij2})\}^{-1}.$$

For the random part z_u is the vector of ones so that $z_u^* = \{\pi'_{ij}\}$.

For the fixed part we have $x^*_{ij1} = x_{ij1}\pi'_{ij}$, $x^*_{ij2} = x_{ij2}\pi'_{ij}$. Suitable starting values can be obtained by first fitting an empirical logit model to the observed proportions

$$\text{logit}(p_{ij}) = \beta_{j0}+\beta_1 x_{ij1}+\beta_2 x_{ij2}+e_{ij}, \qquad (8)$$

where the e_{ij} are regarded as independent level 1 random variables with variances as given below. We can subtract or add say, 0·25 for cell numerators which are equal to the denominators or are zero. In practice just one iteration for (8) is required to obtain suitable starting values for (7).

If we assume that the p_{ij} have a binomial distribution then at each iteration the variances of the p_{ij} are fixed at

$$\pi_i(t)\{1-\pi_i(t)\}n_{ij}^{-1} = \sigma_i^2(t),$$

where n_{ij} is the cell number and π_i is the population mean proportion for the ith cell, with current sample estimates being substituted. There are no covariances between the p_{ij}. As in § 2 the weight matrix is given by

$$V_2 = V_1 + z_u^* z_u^{*\mathrm{T}}\sigma_u^2, \quad V_1 = \text{diag}\{\sigma_i^2(t)\}.$$

This model yields the same numerical estimates as proposed by Longford (1988), who uses a scoring algorithm, on the few comparisons so far carried out. Longford considers a model which results from a Taylor series expansion similar to that used here. He obtains a quasilikelihood model conditional on given values for the level 2 random variables and obtains an unconditional model by assuming a multivariate normal distribution for these variables.

As suggested by Goldstein (1987) we can relax the binomial assumption and simply require the variances to be inversely proportional to n_{ij}. This essentially involves defining dummy explanatory variables $z_{eij} = (n_{ij})^{-\frac{1}{2}}$ which are mutually uncorrelated, so that the level 1 variance for cell (i,j) is, say, $\sigma_{ei}^2 n_{ij}^{-1}$.

To form a comparison with the binomial variance assumption we can define

$$z^*_{eij} = \{\pi_i(1-\pi_i)/n_{ij}\}^{\frac{1}{2}},$$

with variance $\sigma^{*2}_{ei} n_{ij}^{-1}$ and test for $\sigma^{*2}_{ei} = 1$.

Table 1 shows the results of fitting this model. In brackets are the equivalent results from fitting a model assuming a binomial distribution, that is constraining $\sigma^{*2}_{ei} = 1$. The fixed effect estimates show a higher probability of employment for those with a qualification and, to a lesser and non statistically significant extent, for males. The level 1 scale factors are fairly close to 1·0 and the level 2 variance estimate and the estimates for the fixed coefficients are close to those obtained assuming level 1 binomial variation. If a common scale factor for all four categories is fitted the estimate for it is 1·03.

Table 1. *Proportion of individuals employed related to gender and qualification with the intercept varying between areas*

Parameter	Estimate	St. error
Fixed		
Intercept	0·522 (0·526)	
Gender	0·148 (0·149)	0·11 (0·11)
Qualification	1·003 (0·998)	0·11 (0·11)
Random		
Level 2:		
σ^2_u	0·225 (0·234)	0·08
Level 1:		
σ^{*2}_{e1}	1·20	0·20
σ^{*2}_{e2}	0·94	0·16
σ^{*2}_{e3}	0·88	0·14
σ^{*2}_{e4}	1·09	0·16

Values in parentheses are estimates obtained assuming binomial variation with $\sigma^{*2}_{ei} = 1$.

Level 1 parameters ordered by gender within qualification. Gender coded 0 = female, 1 = male; qualification coded 0 = unqualified, 1 = qualified.

Number of level 1 units (cells) = 401; number of level 2 units (areas) = 122.

Table 2 shows the results of allowing the gender coefficient to vary randomly over level 2 units. It is almost uncorrelated with the intercept term, and the between area variance estimate for males, 0·49, is three times as large as that for females, 0·16, although the estimated standard error for the gender difference variance, σ^2_{u1}, is relatively large. The level 1 scale factors deviate from 1·0 rather more than in the previous analysis, so that the estimates based upon the binomial assumption are also more discrepant.

Just under half the cells in the analysis have denominators which are one. For these cells the binomial assumption will be true, so that a test for extra-binomial variation should exclude these cells. When this is done the analysis gives scale factors which are considerably smaller than one, being respectively 0·76, 0·65, 0·53 and 0·57. One explanation for this is that the probability of employment varies across individuals within areas and is related to explanatory factors omitted from the model. A further discussion of such extra-binomial variation models is given by H. Goldstein et al. in an unpublished report.

Table 2. *Proportion of individuals employed related to gender and qualification with the intercept and gender coefficients varying between areas*

Parameter	Estimate	St. error
Fixed		
Intercept	0·511 (0·514)	
Gender	0·162 (0·159)	0·12 (0·12)
Qualification	1·008 (1·005)	0·10 (0·11)
Random		
Level 2:		
σ^2_{u0}	0·158 (0·177)	0·11
σ^2_{u1}	0·325 (0·407)	0·20
σ_{u01}	0·005 (0·116)	0·12
Level 1:		
σ^{*2}_{e1}	1·22	0·21
σ^{*2}_{e2}	0·88	0·15
σ^{*2}_{e3}	0·73	0·15
σ^{*2}_{e4}	0·95	0·16

Values in parentheses are estimates obtained assuming binomial variation with $\sigma^{*2}_{ei} = 1$.

Level 1 parameters ordered by gender within qualification. Gender coded 0 = female, 1 = male; qualification coded 0 = unqualified, 1 = qualified.

The subscripts (0, 1) for level 2 random parameters refer to intercept and gender coefficients respectively.

Number of level 1 units (cells) = 401; number of level 2 units (areas) = 122.

6. QUADRATIC APPROXIMATION

The adequacy of the first order approximation which linearizes the random part of the model needs to be studied further. Considering the random variation at level 2 for illustration, the typical second order term in the Taylor expansion is given by $u^2_k z^2_{uk} f''(H_t)$. The random part now involves random variables and their squares. Assuming normality, we can write expressions for the third and fourth moments of the u_k as follows:

$$E(u_j) = E(u_j u_k^2) = 0, \quad E(u_j^2) = \sigma_j^2, \quad E(u_j u_k) = \sigma_{jk},$$

$$E(u_j^4) = 3\sigma_j^4, \quad E(u_j^2 u_k^2) = 2\sigma_{jk}^2 + \sigma_j^2 \sigma_k^2.$$

This model can be analyzed as follows. The squared random variables are treated as additional random terms whose variances and covariances are not estimated directly but calculated from the current variance and covariance estimates of the basic random variables. The ML3 software for analyzing multilevel data (Rasbash, Prosser & Goldstein, 1989) allows such a specification. In the present case the second derivative is given by

$$f''(H) = f'(H)\{1 - \exp(H)\}\{1 + \exp(H)\}^{-1}.$$

The model in Table 1 has been reanalyzed with a second order approximation, with the random parameter estimates given in Table 1 as starting values. The value of the level 2 variance is slightly reduced, from 0·225 to 0·199, but the other parameter estimates are affected only slightly.

7. Discussion

The general approach adopted in this paper to the estimation of multilevel nonlinear models has the merit that it directly generalizes the linear model case. Estimates of 'shrunken' residuals can also be obtained in the same manner as in linear multilevel models (Goldstein, 1987). In the discrete response data examples analyzed, the approach appears to work well and can be viewed as a straightforward generalization of iteratively reweighted least squares estimation in the single level case.

The addition of a quadratic term to the Taylor expansion of the random component does not appear markedly to change the parameter estimates, although more experience with other data sets would be useful. We can introduce cubic and higher order terms if it is suspected that the linear or quadratic approximations are inadequate. This paper has not pursued the general problems associated with nonlinear estimation which relate to the linear approximation for the fixed part of the model, especially those concerned with convergence properties, and more work is needed in that area. Further work also needs to be carried out on the properties of the estimates of the random parameters, where the current procedure uses a weight matrix based upon multivariate normality assumptions.

Acknowledgements

I am most grateful to Michael Healy, Lindsay Paterson, Ian Plewis, Bob Prosser and Jon Rasbash for their helpful comments, to Cathy Garner for allowing me to use the unemployment data and to two referees. The work was partly supported by a grant from the Economic and Social Research Council.

References

Del Pino, G. (1989). The unifying role of iterative generalized least squares in statistical algorithms. *Statist. Sci.* **4**, 394–408.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* **73**, 43–56.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research.* London: Griffin. New York: Oxford University Press.

Goldstein, H. (1989). Restricted unbiassed iterative generalised least squares estimation. *Biometrika* **76**, 622–3.

Jenss, R. M. & Bayley, N. (1937). A mathematical method for studying the growth of a child. *Human Biol.* **9**, 556–63.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika* **74**, 812–27.

Longford, N. T. (1988). A quasi-likelihood adaptation for variance component analysis. In *Proc. Sect. Comp. Statist., Am. Statist. Assoc.*, pp. 137–42.

McCullagh, P. & Nelder, J. A. (1983). *Generalised Linear Models.* London: Chapman and Hall,

Nelder, J. A. & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221–32.

Rasbash, J., Prosser, R. & Goldstein, H. (1989). *ML2 and ML3: Software for Two-Level and Three-Level Analysis. Users Guide.* London: Institute of Education.

Raudenbush, S. W. & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociol. Educ.* **59**, 1–17.

[*Received October* 1989. *Revised June* 1990]