Improved approximations for multilevel models with binary responses

(Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, A.* 159: 505-13.)

by

Harvey Goldstein

and

Jon Rasbash

Institute of Education

University of London

email: hgoldstn@ioe.ac.uk

Abstract

This paper discusses the use of improved approximations for the estimation of generalised linear multilevel models where the response is a proportion. Simulation studies by Rodriguez and Goldman (1995) have shown that in extreme situations large biases can occur, most notably in the case when the response is binary, the number of level 1 units per level 2 unit is small and the underlying random parameter values are large. An improved approximation is introduced which largely eliminates the biases in the situation described by Rodriguez and Goldman.

Keywords

Binary response, generalised linear model, hierarchical data, marginal model, multilevel model, quasilikelihood, unit specific model

Introduction

Rodriguez and Goldman (1995) point out that existing approximate procedures for estimating parameters of generalised linear multilevel models, in particular those with binary responses, can be seriously biased when the underlying random parameter values are large. These authors use a set of simulated data with the computer programs VARCL (Longford, 1988) and ML3 (Prosser et al, 1991) to demonstrate the extent of these biases. This note describes a procedure which shows a considerable improvement in estimation and is implemented in currently available software. This work was stimulated by the work of Rodriguez and Goldman and we are most grateful to them for helpful discussions and for supplying us with one of their simulated data sets. We now briefly outline the existing procedure and then describe the extensions.

A 2-level binary response model

A simple model which captures the essence of the problem is one where level 1 units, for example mothers, are nested within level 2 units, for example communities. The procedure we describe, however, can be used for any number of hierarchical levels and random coefficients at these levels. For each mother we have a binary response, for example whether or not they received adequate prenatal care during a pregnancy and a set of explanatory variables, measured at either the individual or community level. We write a logit link function

$$\pi_{ij} = f\{X_{ij}\beta + u_j\} = \{1+\exp(-[X_{ij}\beta + u_j])\}^{-1} \qquad (1)$$

for the probability that the *i-th* individual within the *j-th* community received adequate prenatal care. The term $X_{ij}\beta$ is the *ij-th* row of the component of the linear predictor which has fixed coefficients, and u_j represents the random departure for the *j-th* community with $u_j \sim N(0,\sigma_u^2)$. The response y_{ij} for an individual is binary and we make the usual assumption of independent $y_{ij} \sim Bin(1,\pi_{ij})$. In a more general model some of the coefficients β may also vary across level 2 units and the following exposition extends straightforwardly to that case, as it does to models with more than two levels of nesting and with a response which is a proportion.

Our basic approach to estimating the parameters of (1) is first to linearise the exponential function so that it assumes the form of a standard 2-level Normal model and then to apply quasilikelihood estimation using the binomial distribution assumption to define the level 1 variation. Full details can be found in Goldstein (1991) and Rodriguez and Goldman (1995).

We use a first order Taylor expansion for the fixed part about the current estimates. For the second order expansion for the random part we expand about zero, and we show below how this is modified to obtain improved estimates. We obtain at the (t+1)

th iteration of the iterative generalised least squares (IGLS) algorithm (Goldstein, 1986)

$$f(H_{t+1}) = f(H_t) + X_{ij}(\hat{\beta}_{t+1} - \hat{\beta}_t) f'(H_t) + u_j f'(H_t) + u_j^2 f''(H_t)/2 \quad (2)$$

where

$$f'(H) = f(H)[1+\exp(H)]^{-1}, \quad f''(H) = f'(H)[1-\exp(H)][1+\exp(H)]^{-1}$$

The term on the right hand side of the first line of (2) updates the fixed part of the model and in the special case of a single level model provides the updating function and is equivalent to the standard iteratively reweighted least squares algorithm which leads to maximum likelihood estimates. The first term on the second line of (2) is the one suggested by Goldstein (1991) and leads to the first order adjustment which is used in the software packages VARCL and ML3 and by Rodriguez and Goldman (1995). The second term provides a further adjustment which is the basis of the present paper. We note that (2) is essentially a linear model so that procedures for linear multilevel model estimation can be used.

There are two choices we can make for H_t, namely

a) $H_t = X_{ij}\hat{\beta}_t$

b) $H_t = X_{ij}\hat{\beta}_t + \hat{u}_{t,j}$

Choice a) uses only the fixed part predictor for the Taylor expansion and so is referred to as marginal quasilikelihood (MQL) by Breslow and Clayton (1993). Choice b) uses the Taylor expansion about the current estimated residuals, or posterior means, $\hat{u}_{t,j}$, that is conditioning on these for each level 2 unit and is referred to as penalised quasilikelihood (PQL) by Breslow and Clayton, or predictive quasilikelihood since it uses the predicted residual values. Rodriguez and Goldman also consider MQL with a second order correction and show that this improves the estimates, but only slightly. In the remainder of this paper we shall use PQL with the second order term in (2). With the choice b) we expand about $\hat{u}_{t,j}$ for the random part of the model so that the second line of (2) becomes

$$r_u = (u_j - \hat{u}_j) f'(H_t) + (u_j - \hat{u}_j)^2 f''(H_t)/2 \quad (3)$$

and the expansion for the random part is about the current estimate of the level 2 residual rather than zero. For large values of the u_j this will be expected to provide a better linear approximation. In Appendix 1 we show how the estimation for this model is carried out.

Marginal, Population Average and Unit Specific models

Zeger et al (1988) make a distinction between two kinds of models for hierarchically structured data where there is a non-identity link function such as the logit or log.

The model of the present paper is referred to by them as a 'subject specific' model which derives from their consideration of a repeated measures model where 'subject' is level 2. A more general description is 'unit specific' (US) which we shall adopt. Because terms u_j for the higher level units are explicitly included it leads to a specific covariance structure for the responses. An alternative specification is to write what is termed a 'population average' (PA) or 'marginal' (Diggle et al., 1994) model as

$$\pi_{ij} = \{1+\exp(-[X_{ij}\beta^*])\}^{-1} \qquad (4)$$

$$\mathrm{var}(y_{ij}) = \alpha\, \pi_{ij}(1-\pi_{ij})$$

$$\mathrm{cov}(y_j) = V$$

where V can assume particular or general structures, for example an equicorrelation structure. Specifically, it is not an explicit function of the covariance matrix of the random coefficients, although its form is sometimes derived from considering a particular US model and integrating over the random coefficients to obtain the marginal distribution (see for example Bock and Aitkin, 1981).

The two models in general will differ in their covariance structures and hence will provide differing estimates of the fixed coefficients for the same data. The PA model provides no specific information about higher level variation and is therefore useful only for making inferences about average population effects. Thus, (4) allows us directly to estimate the change in response probability corresponding to a unit change in x_{ij} whereas in (1) a unit change in x_{ij} allows us to estimate a change in the response probability *for any given level 2 unit*. Since the link function is nonlinear, this change will depend on u_j.

If we wish to use model (1) to estimate the average population change in probability for a unit change in x_{ij} we can either use an approximation based on the Normality assumption (Zeger et al., 1988) or simulate from the fitted model. In the latter case we would generate a sample of N u_j's assuming Normality, and apply the antilogit transformation to each one for each relevant value of x_{ij}. These transformed values on the probability scale are then averaged to give an estimate of the population mean for the given x_{ij}. By increasing the value of N we can approximate the population mean as accurately as desired (Goldstein, 1995, Chapter 5).

One of the suggested advantages of PA models is the direct estimation of population effects on the probability scale. In view of the fact that these effects are readily estimated from US models this advantage seems negligible. On the other hand, the disadvantage of not being able to provide estimates for higher level structure variation

seems in general to be a major disadvantage of PA models. If there really is a hierarchical structure it seems natural to incorporate it into the model directly. In this sense PA models are not multilevel models at all since there is no explicit hierarchical structure specified. For this reason we do not consider them further here.

Results

In Table 1 we compare first order MQL estimates with second order PQL estimates for the 25 of the simulated data sets used by Rodriguez and Goldman (1995). We have used only the first 25 data sets of Rodriguez and Goldman (1995) since a preliminary study indicated that these provided sufficient accuracy for estimating the bias. We have chosen the most extreme case where the first order MQL estimates perform worst, namely for a three level variance components model with both the level 2 and level 3 variances set to 1. The model from which the data are simulated is

$$\begin{aligned} &\log it(\pi_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + u_k \\ &u_{jk} \sim N(0, \sigma^2_{u2}), \quad u_k \sim N(0, \sigma^2_{u3}) \end{aligned} \tag{7}$$

where i,j,k respectively index the level 1, 2 and 3 units and the true values are given in Table 1. Each data set consists of 2449 level 1 units, 1558 level 2 units and 161 level 3 units with a binary (0,1) response. We have used restricted iterative generalised least squares (Goldstein, 1989) which in the Normal response case is equivalent to REML and have incorporated the adjustment to the variance estimates of the residuals (see Appendix 2). We have used a stringent convergence criterion; namely that for all the parameter estimates the relative change from one iteration to the next is at most 0.001.

Table 1. Mean values of multilevel logit estimates for first 25 simulated data sets used by Rodriguez and Goldman. Column A fits the MQL first order model and column B the second order PQL model. Standard errors of the means are in brackets

	A	B
Parameter (true value)	**MQL first order**	**PQL second order**
Fixed:		
β_0 (0.665)	0.48 (0.03)	0.62 (0.03)
β_1 (1.0)	0.76 (0.03)	0.96 (0.04)
β_2 (1.0)	0.76 (0.01)	0.96 (0.02)
β_3 (1.0)	0.74 (0.03)	0.96 (0.04)
Random:		
σ_{u2} (1.0)	0.09 (0.03)	0.73 (0.02)
σ_{u3} (1.0)	0.73 (0.01)	0.93 (0.02)

It is clear that the second order PQL estimates are a considerable improvement, especially for the level 2 standard deviation and the fixed parameter estimates are close to their true values.

In Table 2 we have carried out a further 200 simulations for the same underlying true model, fitting the first order MQL and PQL models as well as the second order PQL model.

Table 2. Mean values of multilevel logit estimates for 200 simulated data sets using Mln for model given by (7). Column A fits the MQL first order estimates, column B the PQL first order estimates and column C the PQL second order estimates. Standard errors of the means are in brackets

	A	B	C
Parameter (true value)	**MQL first order**	**PQL first order**	**PQL second order**
Fixed:			
β_0 (0.665)	0.512 (0.010)	0.548 (0.011)	0.660 (0.014)
β_1 (1.0)	0.738 (0.012)	0.795 (0.013)	0.965 (0.015)
β_2 (1.0)	0.745 (0.006)	0.805 (0.006)	0.968 (0.008)
β_3 (1.0)	0.767 (0.014)	0.837 (0.015)	1.002 (0.019)
Random:			
σ_{u2} (1.0)	0.119 (0.010)	0.457 (0.006)	0.802 (0.011)
σ_{u3} (1.0)	0.748 (0.004)	0.800 (0.005)	0.968 (0.007)
Percentage of zero estimates at level 2.	54%	9%	0%

The results of these 200 simulations confirm that the only serious bias for the second order PQL estimates is in the level 2 standard deviation, of the order of 20% underestimation. Apart from the level 2 standard deviation parameter, the greatest improvement is in moving from the first order PQL to the second order PQL estimates and both the PQL procedures eliminate most or all of the zero estimates for the level 2 standard deviation. As in the Rodriguez and Goldman study, the standard error estimates for all the parameters, for all estimation methods, are almost unbiased.

In a separate study Ayis (1995) has carried out a comparison of the second order PQL procedure with full maximum likelihood estimation, for a 2-level model with level 2 variances up to the value of 1.0 and with between 24 and 96 level 1 units per level 2 unit. Her study confirms that the second order PQL procedure produces almost unbiased estimates for the fixed parameters and estimates with biases no greater than 4% for the random parameters.

Discussion

We have demonstrated that in the situation considered by Rodriguez and Goldman the second order PQL procedure considerably improves the model estimates, with the greatest improvement occurring with the move from first to second order PQL. Although we have not given details, as Rodriguez and Goldman (1995) demonstrate, a

second order MQL procedure produces only a modest improvement over a first order MQL procedure.

The example chosen is based upon large underlying random parameter values. In the more common case where variances in a variance components model do not exceed about 0.5 the first order PQL model can be expected to perform well, and for smaller variances the first order MQL model will often be adequate. It is also possible that in some circumstances the second order procedure could give worse estimates than the first order one. To establish this would require extensive further simulations which have not yet been undertaken. Likewise, the dependence of the bias on the number of level 1 units within each level 2 unit and the ratio of the number of level 1 to level 2 units requires further study. It does seem, however, that the bias for binary data arises principally from the relatively small number of level 1 units per level 2 unit.

As an analysis strategy, a first order model can be fitted followed by a second order one and note taken of the changes in the estimates. The program *MLn* (Rasbash and Woodhouse, 1995), which is the successor to ML3, has been used for all these calculations. The procedures described here have been applied to handle other link functions and distributions, such as the log-Poisson and logistic-multinomial models.

The principal advantage of the estimation procedures described here is that even for large data sets and numbers of parameters, the computational burden is modest. Full maximum likelihood involving numerical integration is feasible for simple models but becomes intractable when the number of random parameters is moderately large. Gibbs Sampling is another alternative, but is also computationally intensive. The present procedures can be combined with bootstrapping for a final stage of bias reduction. A standard application of a parametric bootstrap (Efron and Tibshirani, 1993) will not yield sarisfactory estimates of the bias, but Kuk (1995) describes an iterated version of the bootstrap which does give asymptotically unbiased estimates, although again computationally intensive.

It would be possible to improve further the approximation given by (2) by considering subsequent terms in the Taylor expansion. For example, if we include a third order term we obtain

$$r_u = (u_j - \hat{u}_j) f'(H_t) + (u_j - \hat{u}_j)^2 f''(H_t)/2 + (u_j - \hat{u}_j)^3 f'''(H_t)/6$$

$$f'''(H_t) = f''(H_t)[1-\exp(H_t)][1+\exp(H_t)]^{-1} - 2f'(H_t)\exp(H_t)[1+\exp(H_t)]^{-2}$$

which will lead to further offsets when estimating the random parameters. We can derive a similar expression for the fourth order term which additionally involves an offset in the fixed part of the model. When these modifications have been implemented, however, there has been little improvement.

Finally, Breslow and Lin (1995) have proposed an alternative approximating approach, but restricted to the 2 level variance components case. We have not compared that approach with the one described in the present paper.

Acknowledgements

We are most grateful to German Rodriguez for his comments, for stimulating the developments discussed in this paper and for supplying a copy of some of his simulated datasets. We are also most grateful to the referees who made helpful comments. This work was carried out under a grant from the Economic and Social Research Council in its Analysis of Large and Complex Datasets programme.

Appendix 1
Estimation using second order adjustments.

Referring to equation (3) and assuming Normality we have, omitting subscripts

$$E(r_u) = \sigma^2_{\hat{u}} f''(H_t)/2,$$
$$\text{var}(r_u) = \sigma^2_{\hat{u}}[f'(H_t)]^2 + \sigma^4_{\hat{u}}[f''(H_t)]^2/2$$
where (A1.1)
$$\sigma^2_{\hat{u}} = \text{var}(u - \hat{u})$$

and the current value of H_t is used. Goldstein (1995, Appendix 2.2) and Waclawiw and Liang (1994) derive formulae for $\sigma^2_{\hat{u}}$.

If we replace the second term in the second line of (2) by its expected value and use (3) we have

$$\begin{aligned}\pi^* &= f(H_{t+1}) + X\hat{\beta}_t f'(H_t) - f(H_t) + \hat{u}_{t,j} f'(H_t) - \sigma^2_{\hat{u}} f''(H_t)/2 \\ &= X\hat{\beta}_{t+1} f'(H_t) + u_j f'(H_t)\end{aligned} \quad \text{(A.1.2)}$$

For the modified response π^* we now have a standard formulation for the second level component of a 2-level model with modified fixed part explanatory variable design matrix $Xf'(H_t)$ and random part explanatory variable $f'(H_t)$. We complete the specification by writing the full model for the observed binary response y_{ij} as

$$y_{ij} = \pi^*_{ij} + e_{ij} z_{eij} = X_{ij}\hat{\beta}_{t+1} f'(H_t) + u_j f'(H_t) + e_{ij} z_{eij}$$
$$z_{eij} = \{\pi_{ij}[1-\pi_{ij}]\}^{\frac{1}{2}} \quad (6)$$
$$E(e_{ij}) = 0. \quad \text{var}(e_{ij}) = 1$$

This definition of the level 1 random variation is based upon the binomial assumption. If we unconstrain $\text{var}(e_{ij})$ then extra-binomial variation models can be fitted. Estimation for (6), with the explanatory variables updated at each iteration, follows the standard procedure as for continuous Normal models, in this case providing quasilikelihood estimates based upon the expected values and the variance function. In each cycle of the IGLS algorithm, the random parameters, the variances and covariances, are first updated and these values used to provide new estimates for the fixed coefficients using generalised least squares. The procedure for updating the random parameters also uses generalised least squares and at this stage the second term in the second line of (4) is used as an offset.

Unless the number of level 2 units is large the estimate of $\sigma^2_{\hat{u}}$ required in (A.1.1) will underestimate the true variance since it takes no account of the sampling variance of the parameters estimates themselves. One, computationally intensive, solution is to carry out a bootstrap estimation at each iteration. Alternatively, we can obtain better estimates using a delta method adjustment which adds a first order or second order correction to the 'naive' estimate. This procedure is described in Appendix 2.

Appendix 2
Delta method estimators for the covariance matrix of residuals

We consider the case of a 2-level Normal model

$$y = X\beta + Z_u u + Z_e e$$

where we require estimates of the level 2 residuals u. Conditional on the observed data and model parameters these are given by (Goldstein, 1995)

$$\hat{u} = \Omega_u Z_u{}^T V^{-1} \tilde{y}, \quad \tilde{y} = y - X\beta \tag{A.2.1}$$

We have, for the comparative variances

$$\text{var}(\hat{u}|y,\beta,\theta) = E[\text{var}(\hat{u}|y,\beta,\theta)] + \text{var}[E(\hat{u}|y,\beta,\theta)] \tag{A.2.2}$$

where the terms on the right hand side of (A.2.2) are regarded as functions of the model parameters and evaluated at the sample estimates. For the *j*-th level 2 unit the first term is given by the usual estimate

$$\begin{aligned} &\Omega_u - R_{(j)}^T V_{(j)}{}^{-1}(V_{(j)} - X(X^T V_{(j)}{}^{-1} X)^{-1} X^T) V_{(j)}{}^{-1} R_{(j)} \\ &R_{(j)} = Z_u^{(j)} \Omega_u \\ &\text{cov}(u) = \Omega_u, \quad V_{(j)} = E(\tilde{y}_{(j)} \tilde{y}_{(j)}^T) \end{aligned} \tag{A.2.3}$$

which adjusts for the sampling variation of the fixed parameter estimates.

We shall use the first order approximation derived from the Taylor expansion about $E(\hat{\theta}) = \theta$, for the covariance matrix of a function, namely

$$\text{cov}[g(\hat{\theta})] \approx \left(\frac{\partial g}{\partial \theta}\right)^T \text{cov}(\hat{\theta}) \left(\frac{\partial g}{\partial \theta}\right)\Bigg|_{\hat{\theta}} \tag{A.2.4}$$

In some circumstances we may wish to have a better approximation, in which case, assuming multivariate Normality, we obtain the additional contribution, evaluated at the sample estimates

$$\frac{1}{4}\left(\frac{\partial^2 g}{\partial\theta^2}\right)^T [2A_1 + A_2]\left(\frac{\partial^2 g}{\partial\theta^2}\right)\Bigg|_{\hat{\theta}}$$

$$A_1 = \{a_{ij}^2\} \quad \text{where } \operatorname{cov}(\hat{\theta}) = \{a_{ij}\}$$
$$A_2 = aa^T \qquad a = \{a_{ii}\}$$

For $\hat{u}$ as a function of the random parameters θ, we have

$$d_k^T = \frac{\partial g}{\partial\theta_k} = [-\Omega_u Z^T V^{-1}\frac{\partial V}{\partial\theta_k}V^{-1} + \frac{\partial\Omega_u}{\partial\theta_k}Z^T V^{-1}]\tilde{y}$$
$$\frac{\partial\Omega_u}{\partial\theta_k} = 0 \;\; \textit{if } \theta_k \textit{ not at level 2.} \qquad \text{(A.2.5)}$$

Note that the elements of $\frac{\partial V}{\partial\theta_k}$ are just the elements of the design vector for the parameter θ_k and that

$$\frac{\partial V^{-1}}{\partial\theta_k} = -V^{-1}\frac{\partial V}{\partial\theta_k}V^{-1}$$

The row vector d_k has q elements, one for each residual at level 2 with $d = \{d_k\}$ an t x r_u matrix where t is the total number of random parameters. The adjustment term in (A.2.2) is therefore

$$d^T \operatorname{cov}(\hat{\theta}) d$$

This procedure for the variance of the estimated residuals is essentially equivalent to that proposed by Kass and Steffey (1989) who give an alternative derivation using the Laplace method. These authors also consider the extra adjustment term based upon the next term in the Taylor expansion as above.

References

Ayis, S.A.M. (1995). *Modelling unobserved heterogeneity: theoretical and practical aspects.* PhD thesis, University of Southampton.

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46, 443-59.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika,* 82, 81-91.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalised linear models. *J. American Statistical Association,* 88, 9-25.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data.* Oxford, Clarendon Press.

Efron, B. and Tibshirani, R.J. (1993). *An introduction to the bootstrap*. London, Chapman and hall.

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

Goldstein, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76, 622-2

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika,* 78, 45-51.

Goldstein, H. (1995). *Multilevel Statistical Models*. London, Edward Arnold; New York, Halstead Press.

Kass,R.E. and Steffey,D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes Models). *J. American Statistical Association,* 84, 717-26.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *J. Royal Statist. Soc.*, B, 57, 395-407.

Longford, N.T. (1988). *VARCL: software for variance component analysis of data with hierarchically nested random effects (maximum likelihood).* Princeton, Educational Testing Service.

Prosser, R., Rasbash, J. And Goldstein, H. (1991). *ML3; software for 3-level analysis: users guide*. London, Institute of Education.

Rasbash, J. and Woodhouse, G. (1995). *MLn Command Reference.* London, Institute of Education.

Rodriguez,G. And Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *J. Royal Statistical Society, A*, 158, 73-90.

Waclawiw, M.A. and Liang, K. (1994). Empirical Bayes estimation and inference for the random effects model with binary response. *Statistics in Medicine*, 13, 541-51

Zeger, S.L., Liang, K. And Albert, P.S. (1988). Models for longitudinal data: a generalised estimating equation approach. *Biometrics*, 44, 1049-60