Improved Approximations for Multilevel Models with Binary Responses

Author(s): Harvey Goldstein and Jon Rasbash

Improved Approximations for Multilevel Models with Binary Responses

By HARVEY GOLDSTEIN† and JON RASBASH

Institute of Education, London, UK

SUMMARY
This paper discusses the use of improved approximations for the estimation of generalized linear multilevel models where the response is a proportion. Simulation studies by Rodríguez and Goldman have shown that in extreme situations large biases can occur, most notably 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 Rodríguez and Goldman.

Keywords: BINARY RESPONSE; GENERALIZED LINEAR MODEL; HIERARCHICAL DATA; MARGINAL MODEL; MULTILEVEL MODEL; QUASI-LIKELIHOOD; UNIT-SPECIFIC MODEL

1. INTRODUCTION

Rodríguez and Goldman (1995) have pointed out that existing approximate procedures for estimating parameters of generalized linear multilevel models, in particular those with binary responses, can be seriously biased when the underlying random parameter values are large. They used 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 paper 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 Rodríguez 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.

2. TWO-LEVEL BINARY RESPONSE MODEL

A simple model which captures the essence of the problem is one where level 1 units, e.g. mothers, are nested within level 2 units, e.g. communities. The procedure that 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, e.g. 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

†*Address for correspondence*: Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK.
E-mail: hgoldstn@ioe.ac.uk

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

for the probability that the ith individual within the jth community received adequate prenatal care. The term $X_{ij}\beta$ is the ijth row of the component of the linear predictor which has fixed coefficients, and u_j represents the random departure for the jth 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 $y_{ij} \sim \mathrm{bin}(1, \pi_{ij})$. 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 equation (1) is first to linearize the exponential function so that it assumes the form of a standard two-level normal model and then to apply quasi-likelihood estimation using the binomial distribution assumption to define the level 1 variation. Full details can be found in Goldstein (1991) and Rodríguez 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 0, and we show below how this is modified to obtain improved estimates. We obtain at the $(t + 1)$th iteration of the iterative generalized least squares (IGLS) algorithm (Goldstein, 1986)

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

where

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

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

There are two choices which we can make for H_t, namely

(a) $H_t = X_{ij}\hat{\beta}_t$ or
(b) $H_t = X_{ij}\hat{\beta}_t + \hat{\mu}_{t,j}$.

Choice (a) uses only the fixed part predictor for the Taylor expansion and so is referred to as marginal quasi-likelihood (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 penalized quasi-likelihood (PQL) by Breslow and Clayton, or predictive quasi-likelihood since it uses the predicted residual values. Rodríguez and Goldman (1995) also considered MQL with a second-order correction and showed that this improves

the estimates, but only slightly. In the remainder of this paper we shall use PQL with the second-order term in equation (2). With choice (b) we expand about $\hat{u}_{t,j}$ for the random part of the model so that the last two terms of equation (2) become

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

and the expansion for the random part is about the current estimate of the level 2 residual rather than about 0. For large values of the u_j this will be expected to provide a better linear approximation. In Appendix A we show how the estimation for this model is carried out.

3.   MARGINAL, POPULATION-AVERAGE AND UNIT-SPECIFIC MODELS

Zeger *et al.* (1988) made a distinction between two kinds of model 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

$$\pi_{ij} = \{1 + \exp(-X_{ij}\beta^*)\}^{-1},\tag{4}$$
$$\mathrm{var}(y_{ij}) = \alpha\pi_{ij}(1 - \pi_{ij}),$$
$$\mathrm{cov}(y_j) = V,$$

where V can assume particular or general structures, e.g. 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, equation (4) allows us directly to estimate the change in response probability corresponding to a unit change in a covariate x_{ij} whereas in equation (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 non-linear, 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_js assuming normality and apply the antilogit transformation to each 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. However, 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 no explicit hierarchical structure is specified. For this reason we do not consider them further here.

4. RESULTS

In Table 1 we compare first-order MQL estimates with second-order PQL estimates for 25 of the simulated data sets used by Rodríguez and Goldman (1995). We have used only the first 25 data sets of Rodríguez 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 the level 3 variances set to 1. The model from which the data are simulated is

$$\text{logit}(\pi_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + u_k, \qquad (5)$$
$$u_{jk} \sim N(0, \sigma_{u2}^2),$$
$$u_k \sim N(0, \sigma_{u3}^2),$$

where i, j and 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 IGLS (Goldstein, 1989) which in the normal response case is equivalent to restricted maximum likelihood and have incorporated the adjustment to the variance estimates of the residuals (see Appendix B). 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.

It is clear that the second-order PQL estimates are a considerable improvement,

TABLE 1
*Mean values of multilevel logit estimates for the first 25 simulated data
sets used by Rodríguez and Goldman (1995)†*

Parameter (true value)	A, MQL first order	B, 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)

†Column A fits the MQL first-order model and column B the second-order PQL model. Standard errors of the means are given in parentheses.

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. 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 0 estimates for the level 2 standard deviation. As in the Rodríguez 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 two-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 that are no greater than 4% for the random parameters.

5. DISCUSSION

We have demonstrated that in the situation considered by Rodríguez 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 Rodríguez and Goldman (1995) demonstrated, a second-order MQL procedure produces only a modest improvement over a first-order MQL procedure.

The example chosen is based on 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

TABLE 2

Mean values of multilevel logit estimates for 200 simulated data sets using MLn for model (7)†

Parameter (true value)	A, MQL first order	B, PQL first order	C, 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)
% 0 estimates at level 2	54	9	0

†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 given in parentheses.

some circumstances the second-order procedure could give worse estimates than the first-order procedure. 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 seems, 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 model 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–multi-nomial 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 satisfactory estimates of bias, but Kuk (1995) describes an iterated version of the bootstrap which does give asymptotically un-biased estimates, although again computationally intensive.

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

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

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

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 two-level variance components case. We have not compared that approach with that described in the present paper.

ACKNOWLEDGEMENTS

APPENDIX A: ESTIMATION USING SECOND-ORDER ADJUSTMENTS

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

$$E(r_u) = \sigma_{\hat{u}}^2 f''(H_t)/2,$$

$$\text{var}(r_u) = \sigma_{\hat{u}}^2 f'(H_t)^2 + \sigma_{\hat{u}}^4 f''(H_t)^2/2 \tag{6}$$

where

$$\sigma_{\hat{u}}^2 = \text{var}(u - \hat{u})$$

and the current value of H_t is used. Goldstein (1995), appendix 2.2, and Waclawiw and Liang (1994) have derived formulae for $\sigma_{\hat{u}}^2$.

If we replace the last term in equation (2) by its expected value and use equation (3) we have

$$\begin{aligned}
\pi^* &= f(H_{t+1}) + X\hat{\beta}_t f'(H_t) - f(H_t) + \hat{u}_{t,j} f'(H_t) - \sigma_{\hat{u}}^2 f''(H_t)/2 \\
&= X\hat{\beta}_{t+1} f'(H_t) + u_j f'(H_t)
\end{aligned} \tag{7}$$

For the modified response π^* we now have a standard formulation for the second-level component of a two-level model with modified fixed part explanatory variable design matrix $X f'(H_t)$ 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

$$y_{ij} = \pi_{ij}^* + e_{ij} z_{eij} = X_{ij}\hat{\beta}_{t+1} f'(H_t) + u_j f'(H_t) + e_{ij} z_{eij}, \tag{8}$$

$$z_{eij} = \{\pi_{ij}(1 - \pi_{ij})\}^{1/2},$$

$$E(e_{ij}) = 0,$$

$$\text{var}(e_{ij}) = 1.$$

This definition of the level 1 random variation is based on the binomial assumption. If we unconstrain $\text{var}(e_{ij})$ then extrabinomial variation models can be fitted. Estimation for equation (8), with the explanatory variables updated at each iteration, follows the standard procedure as for continuous normal models, in this case providing quasi-likelihood estimates based on the expected values and the variance function.

In each cycle of the IGLS algorithm, the random parameters, the variances and co-variances, are first updated and these values used to provide new estimates for the fixed coefficients by using generalized least squares. The procedure for updating the random parameters also uses generalized least squares and at this stage the last term in equation (2) is used as an offset.

Unless the number of level 2 units is large the estimate of $\sigma_{\hat{u}}^2$ required in equation (6) will underestimate the true variance since it takes no account of the sampling variance of the parameter estimates themselves. One, computationally intensive, solution is to carry out a bootstrap estimation at each iteration. Alternatively, we can obtain better estimates by using a delta method adjustment which adds a first-order or second-order correction to the 'naïve' estimate. This procedure is described in Appendix B.

APPENDIX B: DELTA METHOD ESTIMATORS FOR COVARIANCE MATRIX OF RESIDUALS

We consider the case of a two-level normal model

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

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

$$\hat{u} = \Omega_u Z_u^{\mathrm{T}} V^{-1} \tilde{y},$$
$$\tilde{y} = y - X\beta. \tag{9}$$

We have, for the comparative variances,

$$\mathrm{var}(\hat{u}|y, \beta, \theta) = E\{\mathrm{var}(\hat{u}|y, \beta, \theta)\} + \mathrm{var}\{E(\hat{u}|y, \beta, \theta)\} \tag{10}$$

where the terms on the right-hand side of equation (10) are regarded as functions of the model parameters and evaluated at the sample estimates. For the jth level 2 unit the first term is given by the usual estimate

$$\Omega_u - R_{(j)}^{\mathrm{T}} V_{(j)}^{-1} \{V_{(j)} - X(X^{\mathrm{T}} V_{(j)}^{-1} X)^{-1} X^{\mathrm{T}}\} V_{(j)}^{-1} R_{(j)}, \tag{11}$$
$$R_{(j)} = Z_u^{(j)} \Omega_u,$$
$$\mathrm{cov}(u) = \Omega_u,$$
$$V_{(j)} = E(\tilde{y}_{(j)} \tilde{y}_{(j)}^{\mathrm{T}}),$$

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

$$\mathrm{cov}\{g(\hat{\theta})\} \approx \left(\frac{\partial g}{\partial \theta}\right)^{\mathrm{T}} \mathrm{cov}(\hat{\theta}) \left(\frac{\partial g}{\partial \theta}\right)\bigg|_{\hat{\theta}}. \tag{12}$$

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

$$\frac{1}{4} \left(\frac{\partial^2 g}{\partial \theta^2}\right)^{\mathrm{T}} (2A_1 + A_2) \left(\frac{\partial^2 g}{\partial \theta^2}\right)\bigg|_{\hat{\theta}},$$
$$A_1 = \{a_{ij}^2\}$$

where $\mathrm{cov}(\hat{\theta}) = \{a_{ij}\}$

$$A_2 = aa^{\mathrm{T}}$$

and $a = \{a_{ii}\}$.

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

$$d_k^{\mathrm{T}} = \frac{\partial g}{\partial \theta_k} = \left(-\Omega_u Z^{\mathrm{T}} V^{-1} \frac{\partial V}{\partial \theta_k} V^{-1} + \frac{\partial \Omega_u}{\partial \theta_k} Z^{\mathrm{T}} V^{-1}\right) \tilde{y} \tag{13}$$

$$\frac{\partial \Omega_u}{\partial \theta_k} = 0 \qquad \text{if } \theta_k \text{ not at level 2.}$$

The elements of $\partial V / \partial \theta_k$ are just the elements of the design vector for the parameter θ_k and

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

The row vector d_k has q elements, one for each residual at level 2 with $d = \{d_k\}$ a $t \times q$ matrix where t is the total number of random parameters. The adjustment term in equation (10) is therefore

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

This procedure for the variance of the estimated residuals is essentially equivalent to that proposed by Kass and Steffey (1989) who gave an alternative derivation using the Laplace method. They also considered the extra adjustment term based on 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, Southampton.
Bock, R. D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443–459.
Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalised linear models. *J. Am. Statist. Ass.*, **88**, 9–25.
Breslow, N. E. and Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.
Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford: Clarendon.
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.
———(1989) Restricted unbiased iterative generalised least squares estimation. *Biometrika*, **76**, 622–623.
———(1991) Nonlinear multilevel models with an application to discrete response data. *Biometrika*, **78**, 45–51.
———(1995) *Multilevel Statistical Models*. London: Arnold.
Kass, R. E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Statist. Ass.*, **84**, 717–726.
Kuk, A. Y. C. (1995) Asymptotically unbiased estimation in generalized linear models with random effects. *J. R. 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.
Rodríguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *J. R. Statist. Soc.* A, **158**, 73–89.
Waclawiw, M. A. and Liang, K. (1994) Empirical Bayes estimation and inference for the random effects model with binary response. *Statist. Med.*, **13**, 541–551.
Zeger, S. L., Liang, K. and Albert, P. S. (1988) Models for longitudinal data: a generalised estimating equation approach. *Biometrics*, **44**, 1049–1060.