J. R. Statist. Soc. A (2005)
168, *Part* 3, *pp.* 599–613

Variance partitioning in multilevel logistic models that exhibit overdispersion

W. J. Browne,

University of Nottingham, UK

S. V. Subramanian,

Harvard School of Public Health, Boston, USA

K. Jones

University of Bristol, UK

and H. Goldstein

Institute of Education, London, UK

[Received July 2002. Final revision September 2004]

Summary. A common application of multilevel models is to apportion the variance in the response according to the different levels of the data. Whereas partitioning variances is straightforward in models with a continuous response variable with a normal error distribution at each level, the extension of this partitioning to models with binary responses or to proportions or counts is less obvious. We describe methodology due to Goldstein and co-workers for apportioning variance that is attributable to higher levels in multilevel binomial logistic models. This partitioning they referred to as the variance partition coefficient. We consider extending the variance partition coefficient concept to data sets when the response is a proportion and where the binomial assumption may not be appropriate owing to overdispersion in the response variable. Using the literacy data from the 1991 Indian census we estimate simple and complex variance partition coefficients at multiple levels of geography in models with significant overdispersion and thereby establish the relative importance of different geographic levels that influence educational disparities in India.

Keywords: Contextual variation; Illiteracy; India; Multilevel modelling; Multiple spatial levels; Overdispersion; Variance partition coefficient

1. Introduction

Multilevel regression models (Goldstein, 2003; Bryk and Raudenbush, 1992) are increasingly being applied in many areas of quantitative research. Multilevel models reflect the fact that many social and biomedical science data sets contain identifiable units or clusters of observations, e.g. children from the same school or voters from the same electoral ward. Observations from such a data set generally are not independent, and it is important to model any dependence that exists.

By using multilevel modelling, we can account for the interdependence of observations by partitioning the total variance into different components of variation due to various cluster

Address for correspondence: W. J. Browne, School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK.
E-mail: william.browne@nottingham.ac.uk

 0964–1998/05/168599

levels in the data. Partitioning the variance is not simply of technical value; rather the apportioned variances are of substantive interest in much of social science and biomedical research. As we shall demonstrate, ascertaining the relative importance of different levels gives important insights to the level 'at which the action lies'. A by-product of such a decomposition of variation is the intraclass correlation, which is the correlation between two individual observations in the same higher level unit. In the simplest two-level case, the variation can be partitioned into a higher level component based on differences between higher level units and a lower level residual component. These higher level units may be schools, clinics, neighbourhoods or sampling units. Such a model with only one random variable at each level is known as a variance components model. In the two-level variance components model, the intraclass correlation is equal to the percentage of variation that is found at the higher level, which we shall generally call the variance partition coefficient (VPC) although this equivalence is not true for a general multilevel model.

Whereas partitioning variances is straightforward in models with a continuous dependent variable with a normal error distribution at each level, the extension to models with binary responses, or to proportions or counts is less obvious. Goldstein *et al.* (2002) described how to extend the definition to binary response models and gave several methods to evaluate the VPC for such models. In this paper, we describe how to extend the definition to general binomial response models and illustrate three of the methods that were described in Goldstein *et al.* (2002). In particular, we focus on models where the response is a proportion and where the binomial assumption may not be appropriate as overdispersion is likely. Other accounts of partitioning variance in binomial response models can be found in Commenges and Jacqmin (1994) and Snijders and Bosker (1999).

This paper has twin focuses: methodological, in that we wish to present a framework for estimating VPCs for complex models, and substantive, in that we apply this methodology to gain insights on the multiple geographies of illiteracy in India.

The paper is divided up as follows. In Section 2, we provide background material about general binomial response models and discuss two approaches that can be used to account for extra binomial variation. In Section 3, we define the VPC for normal models and suggest how to extend it to binomial models with overdispersion and then describe three methods that we can use for calculating the VPC. Section 4 considers an example application using literacy data from the 1991 Indian census, and we conclude with some discussion on the ideas that are introduced in this paper.

2. Binomial response models

Binary response data occur often in both the social and the biomedical sciences; for example we may have data on whether people have a disease (1) or not (0) or we may have data on whether children pass (1) or fail (0) a particular examination. In the example in Section 4 the response of interest will be whether or not a person is illiterate. Often when such data are collected, although the responses exist at the individual level, the data are not released at the individual level for confidentiality. Consequently, we do not have any predictor variables that are readily available at this level. However, individual data are routinely available merged into a proportion or count at a higher level based on population grouping. In our example, although the data were collected for each individual, they have been transformed into proportions for 12 possible categories of individual within a district in India. As described elsewhere, such data structures are no different from structures where we have individual data (see Subramanian *et al.* (2001)). We shall describe the categories within district units as 'cells' and each person in each cell shares

the same predictor variables, so no additional information is obtained by separating out the observations.

We shall now assume a binomial model for each cell, so that

$$y_i \sim \text{binomial}(n_i, p_i)$$

where, in cell i, y_i is the count of people who are illiterate, n_i is the number of people and p_i is the unknown underlying probability of being illiterate. We shall then fit a relationship between the unknown probability and our predictor variables, typically through a logistic (or alternatively probit) regression. Here for the logistic case we have

$$\text{logit}(p_i) = X_i^{\mathrm{T}}\beta$$

where X_i is a vector of known predictor variables and the expression $X_i^{\mathrm{T}}\beta$ is referred to as the linear predictor (McCullagh and Nelder, 1989). This model assumes independence between the n_i observations and so assuming that we have included the correct variables in the linear predictor then all the variation in the counts, conditional on the estimates of the probabilities, will be binomial with variance equal to $n_i p_i(1 - p_i)$ for cell i.

The binomial is not the only possible distribution for fitting to proportion data and there are other distributions that have greater variation (known as overdispersion) or less variation (known as underdispersion) than the binomial distribution conditional on the values of the p_is.

2.1. Approaches for dealing with overdispersion

There are two main approaches to dealing with the problem of overdispersion. The first method, as described in section 4.5 of McCullagh and Nelder (1989), is the idea of *multiplicative* overdispersion; this idea dates back to Bartlett (1937). Here, we add a multiplicative scale factor to the variance of the response and so we have

$$\text{var}(y_i) = s n_i p_i(1 - p_i)$$

where s is a scale factor which will equal 1 if we have binomial variation, will be greater than 1 if there is overdispersion and less than 1 if there is underdispersion. The advantage of the multiplicative approach is that it allows both overdispersion and underdispersion and we do not need to fit a particular distributional form for the overdispersion. However, we no longer have a true binomial distribution and consequently we cannot write down the likelihood, although moment-based or quasi-likelihood methods exist (see also Williams (1982) for a maximum likelihood approach to a single-level binomial model). We shall, in this paper, use the alternative approach of *additive* overdispersion, which does allow a full likelihood representation. It should be noted that the beta–binomial distribution (Crowder, 1978), which involves assuming that the probabilities for the individual units are themselves from a beta distribution, is an alternative multiplicative approach that does have a likelihood that can be evaluated although the multiplicative factor is in this case not a constant.

In the additive approach we add an additional random term to our model that accounts for the overdispersion. Here, we shall fit a binomial normal model as this links directly with the multilevel approach. Basically we fit for each observation an additional normally distributed error term, which will capture any overdispersion. This model can actually be fitted as a multilevel logistic regression model in standard multilevel modelling software such as MLwiN (Rasbash *et al.*, 2000) with the introduction of an additional 'pseudo'-level. The logistic link model can then be written

$$y_{ij} \sim \text{binomial}(n_{ij}, p_{ij}),$$
$$\text{logit}(p_{ij}) = X_{ij}^{\mathrm{T}}\beta + u_{ij},$$
$$u_{ij} \sim N(0, \sigma_u^2).$$

Here we consider the normally distributed random effects u_{ij} as being at level 2 in a two-level hierarchy which has exactly the same set of units at level 1 and level 2, i.e. each level 2 unit has exactly one level 1 unit. This at first may seem strange, but one could consider the proportion response as consisting of several dichotomous responses and, hence, in essence we are fitting these 0–1 responses at level 1 and the set of responses that make up the proportions comprise the level 2 units, and now we have a standard two-level multilevel model. Of course, we do not actually have to separate the individual responses as the two models are equivalent. This is important, as, when we come to working out the VPCs for the models later, we assume a denominator of 1 at the bottom level. Now that we have established that the additive approach for overdispersion can be fitted simply in a multilevel framework, we shall extend the definition of the VPC to such models.

3. Variance partition coefficients

Goldstein *et al.* (2002) introduced the VPC to describe the percentage of variation in a data set that is attributed to a particular level or classification in the data set. Let us assume a general two-level normal response model as follows:

$$y_{ij} = X_{ij}^{\mathrm{T}}\beta + Z_{ij}^{\mathrm{T}}u_j + e_{ij}, \qquad i = 1, \ldots, n_j, \quad j = 1, \ldots, J,$$
$$u_j \sim \text{MVN}(0, \Omega_u),$$
$$e_{ij} \sim N(0, \sigma_e^2).$$

Here we have β as an $f \times 1$ vector of fixed effects and u_j as an $r \times 1$ vector of random effects for unit j. Let us assume for example that the data are from education and we have children nested within schools; then the VPC_{ij} for the schools is the percentage of variation explained by the school level differences for individual i in school j:

$$\text{VPC}_{ij} = \frac{Z_{ij}^{\mathrm{T}}\Omega_u Z_{ij}}{Z_{ij}^{\mathrm{T}}\Omega_u Z_{ij} + \sigma_e^2}.$$

If we simply have a random intercept at level 2 (i.e. if $r = 1$ and $Z_{ij} = 1\ \forall i, j$) then the VPC is constant across individuals and is equal to the intraclass correlation but, for more complex models where other random terms exist, the VPC will be a function of predictor variables. If we have further levels in the data, then the denominator in the VPC equation will also include the variances at these additional levels.

3.1. Estimation of variance partition coefficients in binomial models with additive overdispersion

Earlier we described how additive overdispersion can be fitted by using a standard multilevel model with an additional pseudo-level, i.e. we assume that the overdispersion terms are at level 2 where level 2 is in fact identical to level 1. To write this out explicitly for a logistic link function, while allowing for simple higher level random effects, we would have

$$\left.\begin{aligned} y_{ijk} &\sim \text{binomial}(n_{ijk}, p_{ijk}),\\ \text{logit}(p_{ijk}) &= X_{ijk}^{\mathrm{T}}\beta + v_k + u_{jk},\\ v_k &\sim N(0, \sigma_v^2),\\ u_{jk} &\sim N(0, \sigma_u^2),\\ i = 1, \ldots, n_{jk}, \quad j = 1, \ldots, n_k, \quad k &= 1, \ldots, K. \end{aligned}\right\} \tag{3.1}$$

Here we now have a three-level model owing to the additional overdispersion level. X_{ijk} are the fixed predictor variables with corresponding coefficients vector β of length f. The u_{jk} are the additive overdispersion random effects with variance σ_u^2. The v_k are the higher level random effects with variance σ_v^2. It should be noted that $n_{jk} = 1 \,\forall\, j, k$ as there is exactly one level 1 unit for each level 2 unit. We shall use a version of this model in our analysis of the illiteracy data that follows. In this case we shall have counts of illiterate people for each of a collection of cells (groups of people) that are nested within districts. In this example i indexes the cells, j is a device to incorporate overdispersion and k indexes the districts with n_k being the number of cells in district k.

We can expand the model by allowing additional higher level random coefficients or allowing the overdispersion to vary with predictor variables as we show later. Now the level 1 variance is binomially distributed and so is a function of the fixed part of the model and Goldstein *et al.* (2002) showed that, therefore, the VPC would be a function of the predictor variables. It should also be noted that in the binomial case with unequal denominators n_{ijk} the variance is also a function of the denominator. Here, however, we shall use the equivalence between the model with proportions and an expanded model with the individual binary responses at a lower level to define the VPC in terms of individuals.

Goldstein *et al.* (2002) considered four approaches that can be used to estimate the VPC in a simple binary response model and we here extend three methods to the overdispersed binomial response case. The fourth method involved treating the response as normally distributed and so is inappropriate here.

3.2. Method A—model linearization

If we consider evaluating p_{ijk} at the mean of the distribution of both the higher level residuals and the overdispersion effects then for the logistic model we have

$$p_{ijk} = \frac{\exp(X_{ijk}^{\mathrm{T}}\beta)}{1 + \exp(X_{ijk}^{\mathrm{T}}\beta)}$$

with first derivative with respect to the fixed part predictor $(X_{ijk}^{\mathrm{T}}\beta)$

$$p'_{ijk} = \frac{p_{ijk}}{1 + \exp(X_{ijk}^{\mathrm{T}}\beta)}.$$

We can then use a first-order Taylor series expansion of p_{ijk} around the above mean to write our model (3.1) in the form

$$\begin{gathered} y_{ijk} = (X_{ijk}^{\mathrm{T}}\beta + v_k + u_{jk})p'_{ijk} + e_{ijk}\sqrt{\{p_{ijk}(1 - p_{ijk})\}},\\ \text{var}(e_{ijk}) = 1. \end{gathered}$$

We then have the formula for VPC_{ijk} at the higher level as follows:

$$\text{VPC}_{ijk} = \frac{\sigma_v^2 p_{ijk}^2/\{1 + \exp(X_{ijk}^{\mathrm{T}}\beta)\}^2}{(\sigma_v^2 + \sigma_u^2)p_{ijk}^2/\{1 + \exp(X_{ijk}^{\mathrm{T}}\beta)\}^2 + p_{ijk}(1 - p_{ijk})}$$

and we estimate this with sample estimates for all the unknown parameters.

3.3. Method B—simulation

The simulation method consists of the following steps.

(a) Fit the three-level model (3.1) by using a suitable estimation method.
(b) From the fitted model simulate a large number (M) of values for the higher level random effects, from the distribution $N(0, \sigma_v^2)$.
(c) For each of the generated higher level random effects, simulate a large number (m) of values for the overdispersion random effects, from the distribution $N(0, \sigma_u^2)$.
(d) For a particular choice of the fixed predictors X_{ijk} compute the $T = Mm$ corresponding values of p_{ijk} ($p_{ijk}^{(r)}$), $r = 1, \ldots, T$, by using the antilogit function. For each of these values compute the level 1 binomial variance $\sigma_{rijk}^2 = p_{ijk}^{(r)}(1 - p_{ijk}^{(r)})$.
(e) For each of the M higher level random-effect draws calculate the mean of the m generated $p_{ijk}^{(r)}$, $p_{ijk}^{(R)}$.
(f) The coefficient VPC_{ijk} is now calculated as

$$\mathrm{VPC}_{ijk} = \sigma_3^2/(\sigma_2^2 + \sigma_1^2)$$

where

$$\sigma_3^2 = \mathrm{var}(p_{ijk}^{(R)}),$$
$$\sigma_2^2 = \mathrm{var}(p_{ijk}^{(r)}),$$
$$\sigma_1^2 = \sum_r \sigma_{rijk}^2/T.$$

3.4. Method C—latent variable approach

Here we assume that the true underlying variable is continuous but we can only observe a binary response that indicates whether the underlying variable is greater or less than a given threshold. In the logistic regression model, the underlying continuous variable will come from a logistic distribution, with a variance of $\pi^2/3$, and hence we substitute this for the level 1 variance, resulting in the formula

$$\mathrm{VPC}_{ijk} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_u^2 + \pi^2/3}.$$

The three methods will not give the same estimates of the VPC although we expect reasonable agreement between methods A and B when the Taylor series approximation is reasonable. Method C may give estimates that are quite different from those of the other methods as it assumes that the level 1 variance is fixed and independent of the predictor variables. We shall see more of the differences between the methods in the example in the next section.

4. An application to the 1991 Indian census data on literacy

Our example consists of data from the 1991 Indian census on levels of literacy in the population that were originally analysed by Subramanian *et al.* (2000, 2001). The data set has two geographical levels—state and district with 442 districts nested within 29 states. Within each district, the individuals are categorized into one of 12 'cells'. Each of the individuals in each cell shares the same values of the three categorical predictor variables: gender, social caste (tribe, caste or general or other) and whether they live in a rural or urban area. The cells are then treated as a lower level unit nested within each district. Some districts do not have individuals from all 12 types of cell and in total we have 5045 cells, which contain between one and

4425806 individuals. The response in each cell is then the proportion of illiterate individuals and we shall fit models with an additive overdispersion term. Our models are then equivalent to fitting individuals nested within cells (as we have no individual level predictors) nested within districts nested within states (in later models) with random effects for the variation at the cell, district and state levels.

All models were estimated by using both the MLwiN (Rasbash *et al.*, 2000) and WinBUGS (Spiegelhalter *et al.*, 2000, 2003) software packages. The three methods that were used to calculate the VPC can take estimates obtained by any estimation method and so we first considered the quasi-likelihood methods that are available in MLwiN. It should be noted that estimates from quasi-likelihood methods such as marginal quasi-likelihood (MQL) and penalized quasi-likelihood (PQL) for fitting binary response models underestimate variance parameters (Rodríguez and Goldman, 1995; Goldstein and Rasbash, 1996). Of the quasi-likelihood methods PQL estimation is less biased than MQL; however, with our data set we had difficulty in achieving convergence with the PQL method (see Goldstein (1991) for details) and so for our models we considered only first-order MQL estimation. It has been shown (Browne, 1998) that Markov chain Monte Carlo (MCMC) methods with diffuse priors are less biased than quasi-likelihood methods for binary response models and we also considered MCMC estimation for our models.

It is important to choose an efficient MCMC estimation method for this data set, particularly because of the large population sizes in some of the individual cells. Initially we tried the single-site Metropolis algorithm that was used in MLwiN (Rasbash *et al.*, 2000) but this gave chains which were heavily correlated and did not converge in a finite number of iterations. MLwiN does not (currently) use the technique of hierarchical centring (Gelfand *et al.*, 1995) which has an enormous effect on the behaviour of the Markov chains in this example. Hierarchical centring involves reparameterizing the model with the aim of using an alternative set of parameters that are less correlated. This will not change the model but is simply an alternative MCMC algorithm for the same model. Without using hierarchical centring we have for each of the cells the underlying probability being estimated by the sum of the fixed effects and several random effects. These terms are extremely highly negatively correlated for cells with large populations and consequently the MCMC methods produce chains that drift rather than converge quickly.

We therefore switched to WinBUGS in which we can implement hierarchically centred formulations of our models. We initially tried WinBUGS 1.3 (Spiegelhalter *et al.*, 2000) but this could not fit our models as the biggest denominators (population sizes) were too big for the adaptive rejection sampler algorithm to work (in the WinBUGS 1.3 implementation, which consequently gives an error message). This problem, however, appears to have been fixed in WinBUGS 1.4 (Spiegelhalter *et al.*, 2003). As will be seen in the first model that follows, the underestimation of the variance parameters in the MQL method results in a large underestimation of the VPC estimates when compared with the MCMC method. This underestimation is common to all the other models that we consider and therefore we quote only MCMC estimates for these other models.

4.1. A simple two-level variance components binary response model with overdispersion

One potential use of the VPC in this data set would be to identify at what level of geography the greatest variability lies. This could be useful if, for example, we wished to initiate a scheme to improve the rate of literacy of the whole population. If we were then to discover that the variation between states is greater than the variation between districts then this will suggest targeting the states with low rates of literacy rather than individual districts.

An alternative use of the VPC is to gain insights into the aetiological aspects that are seen to shape illiteracy. For instance, in India the levels of state and district are not simply administrative units for data collection and dissemination; rather they represent distinct levels at which causal processes affecting illiteracy occur. Whereas a greater variation at the state level would imply dominance of the sociopolitical and financial processes that influence illiteracy, the dominance of the district level would suggest the relative importance of administrative processes. Given that education in India is primarily the responsibility of the states, we can expect a substantial variation between them; at the same time the districts within the states are in charge of implementing educational initiatives and it is not clear how these may differ within the states.

We shall firstly ignore the state level and all predictor variables, however, and fit the model

$$\left.\begin{aligned} y_{ijk} &\sim \text{binomial}(n_{ijk}, p_{ijk}),\\ \text{logit}(p_{ijk}) &= \beta_0 + v_k + u_{jk},\\ v_k &\sim N(0, \sigma_v^2),\\ u_{jk} &\sim N(0, \sigma_u^2),\\ i = 1, \ldots, n_{jk}, \quad j = 1, \ldots, n_k, \quad k &= 1, \ldots, 442. \end{aligned}\right\} \tag{4.1a}$$

Note that this model is a special case of model (3.1) where we have a scalar β_0 instead of a vector β and the X-vector simply consists of a constant. Again i indexes the cells, j is a device to incorporate overdispersion and k indexes the districts, with n_k being the number of cells in district k. We ran first-order MQL estimation and obtained the estimates that are shown in Table 1.

Hierarchical centring then involves reparameterizing the model as follows:

$$\left.\begin{aligned} y_{ijk} &\sim \text{binomial}(n_{ijk}, p_{ijk}),\\ \text{logit}(p_{ijk}) = u^*_{jk} \quad &\text{where} \quad u^*_{jk} = \beta_0 + v_k + u_{jk},\\ u^*_{jk} \sim N(v^*_k, \sigma_u^2) \quad &\text{where} \quad v^*_k = \beta_0 + v_k,\\ v^*_k &\sim N(\beta_0, \sigma_v^2),\\ i = 1, \ldots, n_{jk}, \quad j = 1, \ldots, n_k, \quad k &= 1, \ldots, 442. \end{aligned}\right\} \tag{4.1b}$$

The MCMC algorithm will now consider u^*_{jk} and v^*_k instead of u_{jk} and v_k as parameters in the model that need to be estimated. Then, adding a flat prior for the intercept β_0 and inverse gamma priors for the two variance parameters σ_v^2 and σ_u^2, we can create an 'equivalent' Bayesian formulation of the model. This was run in WinBUGS for 10000 iterations following a burn-in of 500 iterations. The (posterior mean) point estimates and 95% credible intervals are also shown in Table 1. We can clearly see here the underestimation of the MQL method for the variance parameters.

Table 1. First-order MQL parameter estimates and MCMC posterior means (and 95% credible intervals) for model (4.1)

Parameter	*1st-order MQL estimate (standard error)*	*MCMC estimate (credible interval)*
β_0—intercept	−0.021 (0.025)	0.008 (−0.058, 0.076)
σ_v^2—district level variance	0.199 (0.018)	0.383 (0.320, 0.454)
σ_u^2—overdispersion variance	0.764 (0.016)	1.349 (1.293, 1.408)

Table 2. VPC estimates for model (4.1) based on first-order MQL estimates and MCMC posterior means

Method	*District VPC estimate (1st-order MQL) (%)*	*District VPC estimate (MCMC) (%)*
A, linearization	4.01	6.68
B, simulation	3.45 (MCSE 0.016)	5.40 (MCSE 0.033)
C, latent variable	4.68	7.62

The figures in parentheses here and in the other tables of estimates are *estimated asymptotic* standard errors (see Longford (2000)).

We used the three methods to calculate the VPC based on the point estimates from both MQL and MCMC methods. We find that both the linearization and the latent variable approaches are relatively straightforward whereas the simulation approach is slower as it evaluates 25 million estimates of the underlying probability p_{ijk}. As the simulation approach is stochastic we ran it 10 times with different random-number seeds and the Monte Carlo standard error (MCSE) for the average of these runs is included with the estimates here and in later models. The VPC estimates for the three methods are given in Table 2.

We can see that there are differences in the three estimates but all three methods estimate the VPC to be fairly small (between 3% and 5% for MQL). There are, however, bigger differences between the VPC estimates that were produced by using MQL and MCMC estimation, with the VPC estimates from MCMC sampling ranging between 5.4% and 7.7%. For this reason we shall only give estimates from MCMC estimation for the rest of this paper. In Section 4.5 we shall show how to construct interval estimates for the VPC that is produced by methods A and C for this model. It is worth noting that the cell level variation is over three times the variation at the district level and this suggests that within districts there are large differences in literacy rates between the 12 types of people. We could therefore next consider fitting cell level predictor variables to account for these differences.

4.2. Adding cell level predictor variables

We can next consider fitting some fixed predictors into our model. Such a model can be written

$$\left.\begin{aligned} y_{ijk} &\sim \text{binomial}(n_{ijk}, p_{ijk}),\\ \text{logit}(p_{ijk}) &= \beta_0 + X_{ijk}^{\mathrm{T}}\beta + v_k + u_{jk},\\ v_k &\sim N(0, \sigma_v^2),\\ u_{jk} &\sim N(0, \sigma_u^2),\\ i = 1, \ldots, n_{jk}, \quad j = 1, \ldots, n_k, \quad k &= 1, \ldots, 442. \end{aligned}\right\} \tag{4.2a}$$

This model is the same model (3.1) that we used in Section 3 to describe the three methods. We have, however, separated the intercept β_0 from the rest of the fixed effects and so in our example the fixed effects vector β is of length 6. Hierarchical centring can be less effective in models with additional fixed effects as the reparameterization only involves the intercept; however, in the case of overdispersion the use of the pseudo-level means that the predictor variables X_{ijk}^{T} can in fact be written X_{jk}^{T} and we can reparameterize our model as follows:

$$\left.\begin{aligned} y_{ijk} &\sim \text{binomial}(n_{ijk}, p_{ijk}), \\ \text{logit}(p_{ijk}) = u^*_{jk} \quad &\text{where} \quad u^*_{jk} = \beta_0 + X^{\mathrm{T}}_{jk}\beta + v_k + u_{jk}, \\ u^*_{jk} &\sim N(X^{\mathrm{T}}_{jk}\beta + v^*_k, \sigma^2_u), \\ v^*_k &\sim N(\beta_0, \sigma^2_v), \\ i = 1, \ldots, n_{jk}, \quad j = 1, \ldots, n_k, \quad k = 1, \ldots, 442. & \end{aligned}\right\} \tag{4.2b}$$

Again we add flat priors for the fixed effects, β and inverse gamma priors for the two variance parameters σ^2_v and σ^2_u to complete the Bayesian formulation. The reparameterization means that WinBUGS can use conjugate Gibbs sampling for all parameters except u^*_{jk}, which again greatly improves the mixing. The parameter estimates for this model are given in Table 3. Significant main effects for gender, social class and urban or rural habitation were found along with significant interactions between gender and urban or rural habitation, and between caste, social class and urban or rural habitation.

Here we see that accounting for the different types of people has reduced the cell level variation as expected. This model gives different estimated probabilities of illiteracy for the 12 possible cell types and in Table 4 we give estimated probabilities along with estimated VPCs based on methods A and B for the 12 types.

Table 3. Parameter estimates for model (4.2)

Parameter	*MCMC estimate (95% credible interval)*
β_0—intercept	−0.713 (−0.789, −0.639)
β_1—female	1.437 (1.393, 1.481)
β_2—caste	0.696 (0.646, 0.746)
β_3—tribe	0.993 (0.954, 1.033)
β_4—urban	−1.034 (−1.084, −0.983)
β_5—caste × urban	0.350 (0.285, 0.414)
β_6—female × urban	−0.339 (−0.402, −0.276)
σ^2_v—district level variance	0.467 (0.407, 0.539)
σ^2_u—overdispersion variance	0.314 (0.301, 0.328)

Table 4. District VPC estimates for the 12 types of cell in model (4.2)

Type of cell	*Probability of illiteracy*	*VPC method A (%)*	*VPC method B (MCSE) (%)*
Male–other–rural	0.329	8.79	8.05 (0.03)
Male–other–urban	0.148	5.37	5.89 (0.03)
Female–other–rural	0.674	8.77	8.03 (0.04)
Female–other–urban	0.343	8.95	8.13 (0.03)
Male–caste–rural	0.496	9.77	8.57 (0.03)
Male–caste–urban	0.332	8.82	8.06 (0.03)
Female–caste–rural	0.805	6.52	6.69 (0.04)
Female–caste–urban	0.598	9.45	8.41 (0.04)
Male–tribe–rural	0.569	9.61	7.99 (0.03)
Male–tribe–urban	0.320	8.69	7.99 (0.03)
Female–tribe–rural	0.848	5.48	5.97 (0.04)
Female–tribe–urban	0.585	9.53	8.45 (0.04)

Note that the Taylor series approximation method gives larger estimates than does the simulation method for probabilities that are close to 0.5 and smaller estimates for extreme probabilities. By interpolation the methods give the same estimates for probabilities of illiteracy of approximately 0.23 and 0.77. If we were to plot the probability distribution functions for both the standard normal and the logistic distributions then we shall see that they cross at two points and these values, when converted to cumulative probabilities from the logistic distribution, are very close to these probabilities.

Method C is independent of the predictors and hence in this case gives the VPC estimate $0.467/(0.467+0.314+3.29)=11.47\%$ for all types of cell.

The probabilities show that, if we wished to target particular groups of the population to improve literacy, then females from rural areas tend to have the lowest rates of literacy. In terms of variation between districts, the males in rural areas have greater variability in rates of literacy between districts, and so it might also be useful to consider targeting this group in districts where their rates of literacy are low.

4.3. *Adding the state level*

We now consider the additional third level in our data structure: the states in which the districts are nested. We shall firstly consider fitting just a constant term in the fixed part of the model. The extensions to all three methods to include an additional level are fairly routine. We now obtain two VPC estimates, one for each of the state and district level. The simulation-based method is now computationally extremely time consuming as it involves an additional level of looping and hence we reduced the size of M from 5000 to 500, to compare all three methods. We then have the model

$$\left.\begin{aligned} y_{ijkl} &\sim \text{binomial}(n_{ijkl}, p_{ijkl}),\\ \text{logit}(p_{ijkl}) &= \beta_0 + v_l^{(s)} + v_{kl}^{(d)} + u_{jkl},\\ v_l^{(s)} &\sim N(0,\sigma_s^2),\\ v_{kl}^{(d)} &\sim N(0,\sigma_d^2),\\ u_{jkl} &\sim N(0,\sigma_u^2),\\ i=1,\ldots,n_{jkl},\quad j=1,\ldots,n_{kl},\quad k=1,\ldots,n_l,\quad l=1,\ldots,29.& \end{aligned}\right\} \tag{4.3}$$

Here i indexes the cells, j is a device to incorporate overdispersion, k indexes the districts, with n_{kl} being the number of cells in district k, and l indexes the states, with n_l being the number of districts in state l. We have three sets of random effects, the higher level state and district random effects are defined by $v_l^{(s)}$ and $v_{kl}^{(d)}$ respectively and the overdispersion random effects are defined by u_{jkl}. We can again apply hierarchical centring to reparameterize this model and add diffuse prior distributions in a similar way to model (4.1a). Running this model using MCMC estimation gives the estimates that are given in Table 5.

From our initial inspection we can see that the variability between the 29 states is far greater than the variability between the districts within the states. The corresponding VPC estimates are given in Table 6.

This suggests that the variability between the districts that was seen in model (4.1a) is mainly variability between the states in which the districts reside. In terms of targeting population groups this suggests that it would be more useful to consider whole states with poor rates of literacy rather than individual districts.

Table 5. Parameter estimates for model (4.3)

Parameter	*MCMC estimate (95% credible interval)*
β_0—intercept	−0.394 (−0.654, −0.135)
σ^2_s—state level variance	0.455 (0.246, 0.811)
σ^2_d—district level variance	0.042 (0.022, 0.066)
σ^2_u—overdispersion variance	1.346 (1.290, 1.402)

Table 6. VPC estimates for model (4.3)

Method	*State VPC estimate (%)*	*District VPC estimate (%)*
A, linearization	7.59	0.70
B, simulation	6.19 (MCSE 0.08)	0.59 (MCSE 0.001)
C, latent variable	8.87	0.81

4.4. Adding complex variation at the cell level

For our final model we shall consider removing the restriction that the amount of overdispersion is the same for all types of cell. Thus we shall allow a different overdispersion variance for each of the 12 types of cell and we shall reintroduce the fixed effects from model (4.2a) to give

$$\left.\begin{aligned} y_{ijkl} &\sim \text{binomial}(n_{ijkl}, p_{ijkl}),\\ \text{logit}(p_{ijkl}) &= X^{\mathrm{T}}_{ijkl}\beta + v^{(s)}_l + v^{(d)}_{kl} + u_{jkl},\\ v^{(s)}_l &\sim N(0, \sigma^2_s),\\ v^{(d)}_{kl} &\sim N(0, \sigma^2_d),\\ u_{jk_1} &\sim N(0, \sigma^2_{uj}),\\ i = 1, \ldots, n_{jkl}, \quad j = 1, \ldots, n_{kl}, \quad k = 1, \ldots, n_l, \quad l &= 1, \ldots, 29. \end{aligned}\right\} \tag{4.4}$$

Once again we can reparameterize this model in a similar way to model (4.2a) and add diffuse priors. The MCMC parameter estimates for this model are given in Table 7.

Now that we have estimated the overdispersion separately for each type of cell all three methods will give different values of the VPC for each type. As method B is computationally burdensome here we simply quote estimates from methods A and C in Table 8.

Here again we see that the latent variable method always gives higher estimates. The VPC for states varies between 3.6% and 8.5% with method A and 8.5% and 10.1% with method C whereas district differences are less important with VPCs between 1.0% and 2.6% and 2.5% and 3.0% respectively.

4.5. Interval estimates for the variance partition coefficient

It was mentioned earlier that MCMC methods will give less biased estimates for the variance parameters in our data set and hence will give better estimates of the VPC. Another advantage of MCMC methods (and simulation-based estimation methods in general) over the MQL and PQL methods is that they produce not just a point estimate (and standard error) but also chains of sample estimates from the (joint) posterior distribution of interest. It is therefore possible to perform the three VPC methods using the parameter estimates that are obtained at each

Table 7. Parameter estimates for model (4.4)

Parameter	*MCMC estimate (95% credible interval)*
β_0—intercept	−1.055 (−1.306, −0.816)
β_1—female	1.409 (1.365, 1.453)
β_2—caste	0.683 (0.639, 0.728)
β_3—tribe	0.999 (0.951, 1.047)
β_4—urban	−1.025 (−1.070, −0.979)
β_5—caste × urban	0.366 (0.311, 0.420)
β_6—female × urban	−0.319 (−0.375, −0.265)
σ^2_s—state level variance	0.393 (0.213, 0.691)
σ^2_d—district level variance	0.118 (0.100, 0.138)
σ^2_{u1}—male–other–rural	0.195 (0.167, 0.227)
σ^2_{u2}—male–other–urban	0.165 (0.140, 0.194)
σ^2_{u3}—female–other–rural	0.203 (0.175, 0.236)
σ^2_{u4}—female–other–urban	0.094 (0.078, 0.112)
σ^2_{u5}—male–caste–rural	0.141 (0.120, 0.166)
σ^2_{u6}—male–caste–urban	0.120 (0.101, 0.141)
σ^2_{u7}—female–caste–rural	0.466 (0.402, 0.539)
σ^2_{u8}—female–caste–urban	0.125 (0.105, 0.148)
σ^2_{u9}—male–tribe–rural	0.559 (0.476, 0.651)
σ^2_{u10}—male–tribe–urban	0.839 (0.711, 0.986)
σ^2_{u11}—female–tribe–rural	0.795 (0.681, 0.927)
σ^2_{u12}—female–tribe–urban	0.743 (0.634, 0.869)

Table 8. VPC estimates for the 12 types of cell in model (4.4)

Type of cell	*Estimates (%) from the following methods:*			
	Method A—linearization		*Method C—latent variable*	
	VPC state	*VPC district*	*VPC state*	*VPC district*
Male–other–rural	6.63	1.99	9.83	2.94
Male–other–urban	3.64	1.09	9.91	2.97
Female–other–rural	8.12	2.43	9.81	2.94
Female–other–urban	6.93	2.08	10.09	3.02
Male–caste–rural	8.20	2.46	9.97	2.99
Male–caste–urban	6.78	2.03	10.02	3.00
Female–caste–rural	6.39	1.92	9.21	2.76
Female–caste–urban	8.47	2.54	10.01	3.00
Male–tribe–rural	7.74	2.32	9.01	2.70
Male–tribe–urban	5.92	1.77	8.47	2.54
Female–tribe–rural	5.29	1.59	8.55	2.56
Female–tribe–urban	7.48	2.24	8.65	2.59

iteration to construct a chain of VPC estimates from the VPC posterior distribution. This will be too computationally burdensome for the simulation method but not for the other two methods. We shall here again consider the simplest model (4.1a) that we considered previously.

It is now possible to perform the three VPC methods using the parameter estimates obtained at each iteration to construct a chain of VPC estimates from the VPC posterior distribution.

The second method based on simulation (method B) will be computationally prohibitive as the method would need to be repeated at every iteration of the MCMC sampler, whereas the other two methods simply involve plugging in the sampled values of the relevant model parameters at each iteration into the appropriate equations given in Sections 3.2 and 3.4. If we do this for model (4.1a) we obtain (95%) credible interval estimates of (5.64%, 7.82%) from method A and (6.45%, 8.92%) from method C.

5. Discussion

In this paper we have extended the concept of a VPC to the case of proportional data that exhibit overdispersion. We have shown how fitting an additional normally distributed random effect for each proportion can be used to account for overdispersion and how we can calculate the VPC for such a model by using one of three methods. We have seen that the three methods give slightly different results, with the linearization results giving only an approximation to the simulation-based method. The third method based on a latent variable idea is limited in that the VPC that it produces is independent of the fixed predictor variables. This may, however, be attractive if we wish to report results on the underlying 'latent' scale. We have seen how, of the two methods that take account of the individual probability of illiteracy, the simulation method is more accurate as it does not rely on approximations but it becomes too time consuming to calculate for the more complex models with additional levels. We have shown that the choice of estimation procedure is important and that the known underestimation of variance parameters by the MQL procedure will result in an underestimation of the VPC. We have hence used MCMC methods to fit these models and shown how this approach allows us also to calculate interval estimates for the VPC.

We have used as our illustrative example a data set containing the rates of literacy taken from the 1991 Indian census. There are issues with such a large data set about whether

(a) a random-effect model is really necessary owing to the large populations in many cells producing approximately the same estimates as a fixed effect model and
(b) whether the data set should be subsampled to give a smaller data set that is easier to manage.

With regard to substantive conclusions in our illiteracy data set, our results suggest that the bulk of the contextual variation lies between states (as compared with districts). Although this is well known, there has been a tendency to favour research and policy focus on districts given its intuitive appeal as a more proximate geographic level. Indeed, although districts do exhibit variation in illiteracy, their relative variation (after adjusting for the state to which they belong) is rather small. Further reductions in illiteracy, we believe, must come in the form of establishing a politically and financially conducive environment that, in the case of India, is primarily a function of state level processes.

Acknowledgement

WJB is grateful to the Economic and Social Research Council for funding.

References

Bartlett, M. S. (1937) Some examples of statistical methods of research in agriculture and applied biology (with discussion). *J. R. Statist. Soc.*, suppl., **4**, 137–183.
Browne, W. J. (1998) Applying MCMC methods to multilevel models. *PhD Thesis*. Department of Mathematical Sciences, University of Bath, Bath.

Bryk, A. S. and Raudenbush, S. W. (1992) *Hierarchical Linear Models*. Newbury Park: Sage.
Commenges, D. and Jacqmin, H. (1994) The intraclass correlation coefficient: distribution-free definition and test. *Biometrics*, **50**, 517–526.
Crowder, M. J. (1978) Beta-binomial anova for proportions. *Appl. Statist.*, **27**, 34–37.
Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parameterizations for normal linear mixed models. *Biometrika*, **82**, 479–488.
Goldstein, H. (1991) Nonlinear multilevel models with an application to discrete response data. *Biometrika*, **78**, 45–51.
Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn. London: Arnold.
Goldstein, H., Browne, W. J. and Rasbash, J. (2002) Partitioning variation in multilevel models. *Understand. Statist.*, **1**, 223–231.
Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *J. R. Statist. Soc.* A, **159**, 505–513.
Longford, N. T. (2000) On estimating standard errors in multilevel analysis. *Statistician*, **49**, 389–398.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Rasbash, J., Browne, W. J., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. and Lewis, T. (2000) *A User's Guide to MLwin*, 2nd edn. 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.
Snijders, T. and Bosker, R. (1999) *Multilevel Analysis*. London: Sage.
Spiegelhalter, D. J., Thomas, A. and Best, N. G. (2000) *WinBUGS Version 1.3: User Manual*. Cambridge: Medical Research Council Biostatistics Unit.
Spiegelhalter, D. J., Thomas, A., Best, N. G. and Lunn, D. (2003) *WinBUGS Version 1.4: User Manual*. Cambridge: Medical Research Council Biostatistics Unit.
Subramanian, S. V., Duncan, C. and Jones, K. (2000) 'Illiterate people' or 'Illiterate places': the Indian evidence. *Ind. Socl Sci. Rev.*, **2**, 237–274.
Subramanian, S. V., Duncan, C. and Jones, K. (2001) Multilevel perspectives on modelling census data. *Environ. Planng* A, **33**, 399–417.
Williams, D. A. (1982) Extra-binomial variation in logistic linear models. *Appl. Statist.*, **31**, 144–148.