Partitioning Variation in Multilevel Models

Harvey Goldstein, William Browne, and Jon Rasbash

Bedford Group for Lifecourse and Statistical Studies
Institute of Education
London, England

In *multilevel modeling* the residual variation in a response variable is split into component parts that are attributed to various levels. In applied work, much use is made of the percentage of variation that is attributable to the higher level sources of variation. Such a measure, however, makes sense only in simple variance components, Normal response, models where it is often referred to as the *intra-unit correlation*. In this article we describe how similar measures can be found for both more complex random variation in Normal response models and models with discrete responses. In these cases the variance partitions are dependent on predictors associated with the individual observation. We compare several computational techniques to compute the variance partitions.

Keywords: multilevel modeling, variance components models, MQL, PQL, intra-unit correlation, variance partition coefficient

Over the last 15 years, multilevel modeling (see, e.g., Bryk & Raudenbush, 1992; Goldstein, 1995) has been used by researchers in many application areas in both the social and medical sciences. One of the main motivations of multilevel modeling is that generally there are inherently groups of observations within a data set that come from a common source, for example, a school in educational research, or a hospital in medical research. Then two observations chosen randomly from within this particular source are generally not independent, and it is important to model this dependency.

The technique of multilevel modeling accounts for this dependency by partitioning the total variance in the data, having fitted any covariates, into variation due to these sources or *higher level units* and the Level 1 variation that remains. So, for example, in educational research one may consider exam results for schoolchildren, and here one would partition the variation into variation between, and variation within, the higher level units (the schools).

For illustration we consider an educational example of a data set on 4,059 children from 65 schools in the Inner London Education Authority (Goldstein et al., 1993). Here the response variable is the total score for the students in their examinations at age 16 with a predictor being the score on a reading test that each child took at age 11, both test scores having been Normalized.

We consider first the two-level variance components model in which the single reading test predictor x_{1ij} is treated as a fixed effect. The model is then as follows:

$$y_{ij} = \beta_{0j}x_0 + \beta_1 x_{1ij}$$
$$\beta_{0j} = \beta_0 + u_{0j} + e_{0ij}$$
$$x_0 = 1$$
$$var(u_{0j}) = \sigma_{u0}^2 \quad var(e_{0ij}) = \sigma_{e0}^2$$
$$var(y_{ij}|x_{1ij}) = \sigma_{u0}^2 + \sigma_{e0}^2. \qquad (1)$$

A convenient summary of the "importance" of schools is the proportion of the total variance accounted for, which we will call the *variance partition coefficient* (VPC), given by the formula $\tau = \sigma_{u0}^2(\sigma_{u0}^2 + \sigma_{e0}^2)^{-1}$ and, in the variance components model case, this is also a measure of the residual correlation between the responses from two students in the same school; hence the term *intra-unit correlation* is often found with the use of the symbol ρ. Researchers involved in multilevel modeling often find it useful to quote an estimate of the VPC, and discussions of its usefulness and interpretation often occur on the worldwide e-mail discussion list on multilevel models (www.jiscmail.ac.uk/multilevel/). This statistic is also known in the sample survey literature as the *intraclass correlation* and is commonly used as a measure of the extent of clustering.

In this article we show how this summary statistic can be modified for more complex models when there is no simple decomposition of the overall variance and in the case where the response variable is discrete.

Model 1 can be fitted in standard multilevel modeling software, for example, MLwiN (Rasbash et al., 2000), and the results obtained give estimates of .092 for the between-schools variance (σ_{u0}^2) and .563 for the Level 1 variation (σ_{e0}^2). In our example, $\tau = 0.14$ so that there is some clustering effect, and consequently we gain more accurate estimates and confidence intervals by fitting a multilevel model. A 95% confidence interval, using 999 parametric bootstrap replications

(Goldstein, 1995), is (.092, .193). We can also find interval estimates from a Markov Chain Monte Carlo (MCMC) ouput chain (see Conclusions).

GENERAL NORMAL RESPONSE MODELS

The model (Equation 1) is equivalent to fitting parallel lines to each of the 65 schools in the data set. Although the VPC is useful for such a model with a single source of variation at each level, it is less so for a random-coefficient model, for example, a random-slopes regression model where we allow the slopes of the 65 school regressions to vary as follows:

$$y_{ij} = \beta_{0j}x_0 + \beta_{1j}x_{1ij}$$
$$\beta_{0j} = \beta_0 + u_{0j} + e_{0ij}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$
$$x_0 = 1$$
$$var(u_{0j}) = \sigma_{u0}^2, \quad var(u_{1j}) = \sigma_{u1}^2,$$
$$cov(u_{0j}, u_{1j}) = \sigma_{u01}, \quad var(e_{0ij}) = \sigma_{e0}^2. \qquad (2)$$

Here $\tau = (\sigma_{u0}^2 + 2\sigma_{u01}x_{1ij} + \sigma_{u1}^2 x_{1ij}^2)(\sigma_{u0}^2 + 2\sigma_{u01}x_{1ij} + \sigma_{u1}^2 x_{1ij}^2 + \sigma_{e0}^2)^{-1}$.

In this case the VPC is not the same as the intra-unit correlation because for two children with scores x_{1i_1j}, x_{1i_2j}, the correlation is given by

$$\frac{(\sigma_{u0}^2 + \sigma_{u01}(x_{1i_1j} + x_{1i_2j}) + \sigma_{u1}^2 x_{1i_1j}x_{1i_2j})}{\sqrt{(\sigma_{u0}^2 + 2\sigma_{u01}x_{1i_1j} + \sigma_{u1}^2 x_{1i_1j}^2 + \sigma_{e0}^2)(\sigma_{u0}^2 + 2\sigma_{u01}x_{1i_2j} + \sigma_{u1}^2 x_{1i_2j}^2 + \sigma_{e0}^2)}}.$$

More generally, the VPC may be a function of several predictor variables (at either level). Thus its simplicity as a measure of clustering is diminished, although for any combination of predictor variables in the model the VPC can be calculated. Figure 1 plots the VPC as a function of reading test score. Of particular interest here is the way in which the importance of the school attended increases markedly with increasing reading test score above the mean.

DISCRETE RESPONSE MODELS

We now consider a multilevel model with a binary response, but our remarks apply more generally to models for proportions, for different nonidentity link functions,
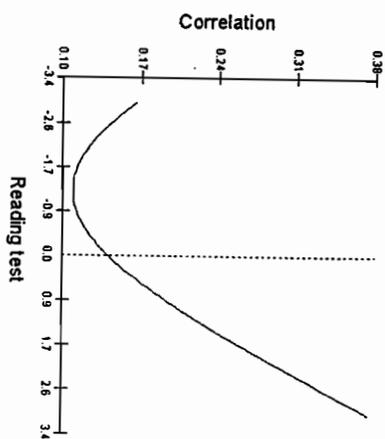
and where the response is a count, in fact to any nonlinear model. For a (0, 1) response, the model that is analogous to Equation 1 is

$$E(y_{ij}) = \pi_{ij} = g(\beta_0 x_0 + \beta_1 x_{1ij} + u_{0j})$$
$$y_{ij} \sim Bernoulli(\pi_{ij})$$
$$u_{0j} \sim N(0, \sigma_{u0}^2). \tag{3}$$

Remembering that the response is just (0, 1), we see that unlike in the Normal case here the Level 1 variance depends on the expected value, $\mathrm{var}(y_{ij}) = \pi_{ij}(1 - \pi_{ij})$ as the fixed predictor in the model depends on the reading test score. Therefore as we are considering a function of the predictor variable x_1, once again a simple VPC is not available. Furthermore, the Level 2 variance, σ_{u0}^2, is measured on the logistic scale and so is not directly comparable to this Level 1 variance.

If we still wish to produce a measure, however, albeit dependent on x_1, the following procedures will provide at least approximate estimates.

Model Linearization (Method A)

Using a first-order Taylor expansion (see, e.g., Goldstein & Rasbash, 1996), we can write Equation 3 in the form

$$y_{ij} = (\beta_0 + \beta_1 x_{1ij})\pi'_{ij} + u_j\pi'_{ij} + e_{ij}\sqrt{\pi_{ij}(1 - \pi_{ij})}$$

where we evaluate π_{ij} at the mean of the distribution of the Level 2 random effect, that is, for the logistic model

$$\mathrm{var}(e_{0ij}) = 1,$$

FIGURE 1   Plot of variance partition coefficient for different reading test scores.

so that, for a given value of x_1, we have

$$\mathrm{var}(y_{ij}|x_{1ij}) = \sigma_{u0}^2 \pi_{ij}^2[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-2} + \pi_{ij}(1 - \pi_{ij})$$

and   (4)

$$\tau = \sigma_{u0}^2 \pi_{ij}^2[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-2}\{\sigma_{u0}^2 \pi_{ij}^2[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-2} + \pi_{ij}(1 - \pi_{ij})\}^{-1},$$

where sample estimates are substituted.

Simulation (Method B)

This method is general and can be applied to any nonlinear model without the need to evaluate an approximating formula. It consists of the following steps:

1. From the fitted model (say, Equation 3) simulate a large number, m (say, 5,000) values for the Level 2 residuals from the distribution $N(0, \sigma_{u0}^2)$, using the sample estimate of the variance.

2. For a particular chosen value(s) of x_1, compute the m corresponding values of $\pi_{ij}(\pi_{ij}^*)$ using Equation 4. For each of these values compute the Level 1 variance $v_{ij} = \pi_{ij}^*(1 - \pi_{ij}^*)$.

3. The coefficient is now estimated as

$$\tau = v_2(v_2 + v_1)^{-1}$$
$$v_2 = \mathrm{var}(\pi_{ij}^*), \quad v_1 = E(v_{ij}).$$

A Binary Linear Model (Method C)

As a very approximate indication for the VPC we can consider treating the (0, 1) response as if it were a Normally distributed variable as in Equation 1 and estimate the VPC as in that case. This will generally be acceptable when the probabilities involved are not extreme, but if any of the underlying probabilities are close to 0 or 1, this model would not be expected to fit well and may predict probabilities outside the (0, 1) range.

A Latent Variable Approach (Method D)

In some circumstances we may wish to think of an observed (0, 1) as arising from an underlying continuous variable so that a 1 is observed when a certain threshold is

exceeded, otherwise a 0 is observed. For the logit model we have the underlying logistic distribution

$$f(x) = \frac{\exp(x)}{[1+\exp(x)]^2},$$ (5)

with cumulative distribution function

$$\int_{-Y}^{\infty} f(x)dx = [1+\exp(-Y)]^{-1}.$$ (6)

The right hand side of Equation 6 is simply the logit link function model with Y as the linear component incorporating Level 2 variation. The variance for the standard logistic distribution in Equation 5 is $\pi^2/3 = 3.29$ so we take this to be the Level 1 variance, and both the Level 1 and Level 2 variances are on a continuous scale. We now simply calculate the ratio of the Level 2 variance to the sum of the Level 1 and Level 2 variances to obtain the VPC. A similar device can be used for the probit link function.

This approach may be reasonable where the (0, 1) response is, say, derived from a truncation of an underlying continuum such as a pass–fail response based on a continuous mark scale but would seem to have less justification when the response is truly discrete, such as mortality or voting (for further discussion, see Snijders & Bosker, 1999, chap. 14).

An Example

We illustrate the procedures using data on voting patterns (Heath, Yang, & Goldstein, 1996).

The model response is whether, in a sample of 800 respondents, they expressed a preference for voting Conservative ($y_{ij} = 1$) in the 1983 British general election. Several covariates were originally fitted, but for simplicity we fit only an intercept model, namely

$$E(y_{ij}) = \pi_{ij} = \exp(\beta_0 + u_{0j})[1+\exp(\beta_0 + u_{0j})]^{-1}$$
$$y_{ij} \sim Bernoulli(\pi_{ij})$$
$$u_{0j} \sim N(0, \sigma_{u0}^2).$$

The fitted model parameters are as follows (using PQL2 with IGLS; Goldstein & Rasbash, 1996):

TABLE 1
Level 2 and Level 1 Estimated Variances With VPC

Variance	Method			
	A	B	C	D
Level 2	.0086	.0083	.0088	0.142
Level 1	.2460	.2380	.2370	3.290
VPC	.0340	.0340	.0360	0.043

Note.   VPC = variance partition coefficient

$$\hat{\beta}_0 = -0.256, \quad \hat{\sigma}_{u0}^2 = 0.142.$$

The corresponding estimates are given in Table 1.

There is, as expected, good agreement for methods A, B, and C, with that for D somewhat larger. Now, however, we choose an extreme value. In chapter 8 of Rasbash et al. (2000) these data are fitted with four predictor variables measuring political attitudes. The scales of these predictor variables are constructed such that lower values correspond to left-wing attitudes. If we take the set of values corresponding to the 5th percentile point of each of the four predictors, then the predicted value on the logit scale is approximately -2.5, corresponding to a low probability of voting Conservative of .076. At this value of the predictor the respective VPCs for Methods A and B are .0096 and .0111, which are again in reasonable agreement. For Method C, fitting the four predictor variables, we obtain a value of .0257, which is very different. For Method D the residual Level 2 variance does not change very much from the previous model, and we obtain a VPC of .047, which is also different to the others. We note that the VPC for Methods C and D do not depend on the value of the linear predictor.

CONCLUSIONS

In summary, if a VPC is required for nonlinear models, then it should be computed for a range of values of the predictor variables. Methods A or B are the most appropriate when one wishes to measure the variability on the probability scale. Method C becomes unreliable for extreme probabilities. The advantage of Method B is that it does not involve any approximation and is simple and fast to compute. One can readily extend the computations for random coefficient models; with Method B this simply involves simulating from a multivariate Normal distribution using the estimated Level 2 covariance matrix. Method D attempts to measure the variability on an underlying continuous logistic (or Normal) scale, and as with continuous re-

sponse variance component models such as Model 1, the VPC does not then depend on the values of the linear predictor. If one wishes to make inferences on such an underlying continuous scale, then Method D is appropriate. The choice of whether to report on the probability scale or on an underlying continuous scale will depend on the application; in the example of this article it seems natural to report directly on the probability scale rather than on an assumed underlying continuous scale of "propensity" to vote Conservative.

We can obtain an interval estimate for any of our estimates via the bootstrap or using the results from an MCMC estimation run. In the former case we would generate a suitable number (e.g., 999) of bootstrap data sets with associated model fits and then, using one of the earlier procedures, obtain a sample of (999) values from which quantiles can be estimated, as in the first example of this article. In the MCMC case we would run a chain (of length, say, 5,000) in which each MCMC cycle in the chain generates a sample from the posterior distribution of the parameters, and using one of the above procedures this likewise will produce a sample of values from the posterior distribution of the chosen VPC (see Rasbash et al., 2000, chap. 15).

REFERENCES

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Arnold.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A, 159,* 505–513.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuall, D., et al. (1993). A multilevel analysis of school examination results. *Oxford Review of Education, 19,* 425–433.

Heath, A., Yang, M., & Goldstein, H. (1996). Multilevel analysis of the changing relationship between class and party in Britain 1964-1992. *Quality and Quantity, 30,* 389–404.

Rasbash, J., Browne, W. J., Goldstein, H., Yang, M., Plewis, I., Healy, M., et al. (2000). *A user's guide to MLwiN* (2nd ed.). London: Institute of Education.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. London: Sage.

APPENDIX

Sample MLwiN Macros for Methods A and B

Methods C and D are trivial to evaluate. Macros follow for evaluating Methods A and B for any two-level binomial model using the MLwiN macro language. The macros require the set of explanatory variable (x) values for which the variance partition coefficient is to be calculated in c151. Column c152 contains the set of x values that have random coefficients in the model. The text for these macros can be executed in an MLwiN macro window. The commands Print b7 or Print b8 display the results from Methods A and B in the output window. For example, for Model 3, if

we require the VPC for a value $x_{ij} = 0.5$ we have C151 = (1, 0.5), C152 = (1), $x_0 \equiv 1$ (defining the intercept).

Macro for Method A

```
note c151 contains values for set of x variables for which
note: we wish to calculate variance partition coefficient
note c152 contains subset of c151 with random effects at level 2
note calculate (XB) and store in b2 and pi=antilogit(XB) in b3
pick 1 c151 b2
calc c153=(~c151)*.c98
note calc. level/2 variance for chosen vals. of expl. Vars: store in b4
calc c153=(~c152)*.omega(2)*.c152
pick 1 c153 b4
note pi^2*Su^2. Su is level 2 variance matrix
calc b3=alog(b2)
calc b5=b3^2*b4
calc b6=b5/(1+expo(b2))^2
calc b7=b6/(b6+b3*(1-b3))
Print b7
```

Macro for Method B

```
note c151 contains values for set of x variables for which
note: we wish to calculate variance partition coefficient
note c152 contains subset of c151 with random effects at level 2
note calculate (XB) and store in b2
calc c153=(~c151)*.c98
pick 1 c153 b2
note calc. level/2 variance for chosen vals. of expl. Vars: store in b4
calc c153= (~c152)*.omega(2)*.c152
pick 1 c153 b4
nran 5000 c154
calc c154=alog(c154*b4^0.5+b2)
aver c154 b1 b3 b2
calc c154=c154*(1-c154)
aver c154 b5 b1
calc b8=b2^2/(b1+b2^2)
Print b8
```