Multilevel Modeling: When and Why[1]

J. Hox

University of Amsterdam & Utrecht University
Amsterdam/Utrecht, the Netherlands

Abstract: Multilevel models have become popular for the analysis of a variety of problems, going beyond the classical individuals-within-groups applications. This chapter gives a brief summary of the reasons for using multilevel models, and provides examples why these reasons are indeed valid reasons. Next, recent (simulation) research is reviewed on the robustness and power of the usual estimation procedures with varying sample sizes.

1 Why multilevel data need multilevel models

Multilevel models are models specifically geared toward the statistical analysis of data that have a hierarchical or clustered structure. Such data arise routinely in various fields, for instance in educational research, where pupils are nested within schools, family studies with children nested within families, medical research with patients nested within physicians or hospitals, and biological research, for instance the analysis of dental anomalies with teeth nested within different persons' mouths. Clustered data may also arise as a result of the specific research design. For instance, in large scale survey research the data collection is usually, for economic reasons, organized in some sort of multistage sampling design that results in clustered or stratified design. Another example are longitudinal designs; one way of viewing longitudinal data is as a series of repeated measurements nested within individual subjects.

Older approaches to the analysis of multilevel data, such as reviewed in Huetnner and Van den Eeden (1982), tend to simply ignore the question, and perform the analysis by disaggregating all data to the lowest level and subsequently applying standard analysis methods. The problems created by this approach were recognized (e.g., Burstein, 1980), but remained statistically intractable.

The magnitude of the statistical problem can be illustrated by a simple example from sample surveys. Survey statisticians have long known that the extend to which the samples are clustered affects the sampling variance. In his classic work, Kish (1965) provides a detailed examination of the effect of cluster sampling on sampling

[1] Pp. 147-154 in: I. Balderjahn, R. Mathar & M. Schader (Eds.). *Classification, data analysis, and data highways*. New York: Springer Verlag.

variance. He defines the **design effect** (*deff*) as the ratio of the operating sampling variance to the sampling variance that applies to simple random sampling. Thus, *deff* is the factor with which the simple random sampling variance must be multiplied to provide the actual operating sampling variance. Kish (1965) describes how *deff* can be estimated for various sampling designs. In simple cluster sampling with equal cluster sizes *deff* can be computed by $deff=(1+rho(n_{clus}-1))$ where rho is the intraclass correlation and n_{clus} is the common cluster size. It is clear that *deff* equals one only when either the intraclass correlation is zero, or the cluster size is one. In all other situations *deff* is larger than one, which implies that standard statistical formulas will underestimate the sampling variance, and therefore lead to significance tests with an inflated alpha level (type I error rate).

Using *deff*, the standard statistical formulas can be adjusted to reflect the true sampling variance. If such adjustments are made, the impact of cluster sampling on the operating alpha level is often rather large. For example, Barcikowski (1981) examines the effect of cluster sampling on the actual alpha level of a *t*-test performed at a nominal alpha level of 0.05. With a small intraclass correlation of rho=0.05 and a cluster size of 10, the operating alpha level is 0.11. With larger intraclass correlations and larger cluster sizes, the operating alpha level increases rapidly. For another example, we can investigate the effect of cluster sampling in educational research. In educational research, data is often collected from classes. Assuming a common class size of 25 pupils, and a typical intraclass correlation for school effects of rho=0.10, we calculate an operating alpha level of 0.29 for tests performed at a nominal alpha level of 0.05! Clearly, in such situations **not** adjusting for clustered data produces totally misleading significance tests.

The examples given above are confirmed by simulation research by Tate and Wongbundhit (1983). They generate multilevel data following a regression model, and conclude that the estimates of the regression coefficients are unbiased, but have a much larger sampling variance than ordinary least squares (OLS) methods would produce. Again, significance tests ignoring the multilevel structure of the data would produce spuriously significant effects.

2 The multilevel regression model

The multilevel regression model is known in the research literature under a variety of names, such as 'random coefficient model' (de Leeuw & Kreft, 1986; Longford, 1993), 'variance component model' (Longford, 1993), and 'hierarchical linear model' (Raudenbush & Bryk, 1986; Bryk & Raudenbush, 1992). It assumes hierarchical data, with one response variable measured at the lowest level and explanatory variables at all existing levels. Conceptually the model is often viewed as a hierarchical system of regression equations. For example, assume we have data in *J* groups or contexts, and a different number of individuals N_j in each group. On the individual (lowest) level we have the dependent variable Y_{ij} and the explanatory variable X_{ij}, and on the group level we have the explanatory variable Z_j. Thus, we have a separate regression equation in each group:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}. \tag{1}$$

The β_j are modeled by explanatory variables at the group level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}, \tag{2}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \tag{3}$$

Substitution of (2) and (3) in (1) gives:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{1j} X_{ij} + u_{0j} + e_{ij} \tag{4}$$

In general there will be more than one explanatory variable at the lowest level and also more than one explanatory variable at the highest level. Assume that we have *P* explanatory variables X at the lowest level, indicated by the subscript p (p=1..P), and *Q* explanatory variables Z at the highest level, indicated by the subscript q (q=1..Q). Then, equation (4) becomes the more general equation:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{pq} Z_{qj} X_{pij} + u_{pj} X_{pij} + u_{0j} + e_{ij} \tag{5}$$

The estimators generally used in multilevel analysis are Maximum Likelihood (ML) estimators, with standard errors estimated from the inverse of the information matrix. These standard errors are used in the Wald test (Wald, 1943); the test statistic *Z = parameter / (st.error param.)* is referred to the standard normal distribution to establish a p-value for the null-hypothesis that in the population that specific parameter is zero.
Computing the Maximum Likelihood estimates requires an *iterative* procedure. At the beginning the computer program generates reasonable starting values for the various parameters, often single level OLS estimates. After one iteration, we have Generalized Least Squares (GLS) estimates. When the iterative process converges, we have ML estimates. GLS estimates require much less computing time, which makes them attractive for computer-intensive procedures like simulation and bootstrapping. Also, it may be possible to obtain GLS estimates, in situations where the iterative procedure does not converge (cf. Goldstein, 1995, p23).
Two different varieties of Maximum Likelihood estimation are commonly used in multilevel regression analysis. The one is Full Maximum Likelihood (FML); in this method both the regression coefficients and the variance components are included in the likelihood function. The other is Restricted Maximum Likelihood (RML); here only the variance components are included in the likelihood function. The difference is that FML treats the estimates for the regression coefficients as known quantities when the variance components are estimated, while RML treats them as estimates

that carry some amount of uncertainty (Bryk and Raudenbush, 1992; Goldstein, 1995). Since RML is more realistic, it should, in theory, lead to better estimates, especially when the number of groups is small (Bryk & Raudenbush, 1992). FML has two advantages over RML: the computations are generally easier, and since the regression coefficients are included in the likelihood function, a likelihood ratio can be used to test for differences between two nested models that differ only in the fixed part (the regression coefficients). With RML only differences in the random part (the variance components) can be tested this way.

The assumptions of are that the residual errors at the lowest level e_{ij} have a normal distribution with a mean of zero and a common variance σ^2 in all groups. The second level residual errors u_{0j} and u_{pj} are assumed to be independent from the lowest level errors e_{ij}, and to have a multivariate normal distribution with means of zero. Other assumptions, identical to the common assumptions of multiple regression analysis, are fixed predictors and linear relationships. The standard errors generated by the ML procedure are asymptotic; meaning we need fairly large samples at all levels.

3 The accuracy of parameter estimates

The assumptions stated above generate questions about the accuracy of the various estimation methods when these assumptions are false. Most research in this direction uses simulation methods, and investigates the accuracy of the fixed and random parameters with small sample sizes and nonnormal data. Comparatively less research investigates the accuracy of the standard errors used to test specific model parameters.

3.1 Accuracy of fixed parameters and their standard errors

The estimates for the regression coefficients appear generally unbiased, for OLS, GLS, as well as ML estimation. OLS estimates have a larger sampling error; Kreft (1996), reanalyzing results from Kim (1990) estimates that they are about 90% efficient.

The OLS based standard errors are known to be severely biased downward, which has been illustrated in the introduction section. The asymptotic Wald tests used in most multilevel software. This assumes either large samples (of groups) or equal group sizes.

The power of the Wald test for the significance of the individual level regression coefficients depends on the total sample size. The power of tests of higher level effects and cross-level interactions depends more strongly on the number of groups than on the total sample size. Both simulations (Van der Leeden & Busing, 1994; Mok, 1995) and analytic work (Snijders & Bosker, 1993) suggest a trade-off between sample sizes at different levels. For accuracy and high power a large number of groups appears more important than a large number of individuals per group. Bryk and Raudenbush (1992) argue that the test statistic should rather be referred to a Student distribution with J-p-1 degrees of freedom (number of groups - number of parameters estimated - 1), citing simulations by Fotiu (1989, cited by Bryk and Raudenbush, 1992). Simulations by Van der Leeden & Busing (1994) and Van der Leeden et al. (1997) suggest that when assumptions of normality and large samples are not met, the ML estimates are still unbiased, but the standard errors are somewhat biased downward. GLS estimates of fixed parameters and their standard errors are somewhat less accurate, but workable.

3.2 Accuracy of random parameters and their standard errors

Estimates of the residual error at the lowest level are generally accurate. The group level variance components are generally underestimated, with FML somewhat more that with RML. GLS variance estimates are less accurate than ML estimates, and for accurate estimates many groups (>100) are needed (Busing, 1993; Van der Leeden & Busing, 1994; Afshartous, 1995).

The asymptotic Wald test for the variance components implies the unrealistic assumption that they are normally distributed. For this reason, other approaches have been advocated, among which estimating the standard error for sigma (the square root of the variance, Longford, 1993), and using the likelihood ratio test. Bryk & Raudenbush (1992) advocate a chi-square test based on the OLS residuals. The literature contains no comparisons between these methods. Simulations by Van der Leeden et al. (1997) show that the standard errors used for the wald test are generally estimated too small, with RML more accurate than FML. Symmetric confidence intervals around the estimated value also do not perform well. Van der Leeden et al. show that the bootstrap is a promising alternative to obtaining more accurate variance estimates, standard errors, and confidence intervals, provided the bootstrapping method samples cases and not residuals.

3.3 Accuracy and sample size

It is clear that with increasing sample sizes at all levels, estimates and their standard errors become more accurate. Kreft (1996) suggests a rule of thumb which she calls the '30/30 rule.' To be on the safe side, researchers should strive for a sample of at least 30 groups with at least 30 individuals per group. From the various simulations reviewed above, this seems sound advice if the interest is mostly in the fixed parameters. For certain applications it may be wise to modify this rule of thumb. Specifically, if there is strong interest in cross-level interactions, the number of groups should be larger, which leads to a 50/20 rule: about fifty groups with about 20 individuals per group. If there is strong interest in the random part, the variance and covariance components, the number of groups should be considerably larger, which leads to a 100/10 rule: about 100 groups with about 10 individuals per group. These rules of thumb take into account that there are costs attached to data collection, so if the number of groups is increased, the number of individuals per group decreases. In some cases this may not be a realistic reflection of costs. For instance, in school research the extra cost will be incurred when an extra class is included. Testing only part of the class instead of all pupils will usually not make much difference in the data collection cost. Given a limited budget, an optimal design should reflect the various costs of data collection. Snijders and Bosker (1994) discuss the problem of choosing sample sizes at two levels while taking costs into account.

4. Analysis of proportions and binary data

Multilevel analysis of proportions is uses a generalized linear models with a logit link, which gives us the model:

$$\text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_{10}X_{ij} + U_{0j} \qquad (6)$$

The observed proportions P_{ij} are assumed to have a binomial distribution with known variance

$$\text{var}(P_{ij}) = (\pi_{ij}\,(1 - \pi_{ij}\,))/n_{ij}$$

In the software, this variance is usually specified by including the predictor s_{ij} = _var(P_{ij}) in the random part, with associated variance constrained to one. The π_{ij} are estimated by prediction from the current model. If the variance term is not constrained to one, but estimated, we can model over- and underdispersion. If the extrabinomial variation is significantly different from one, this is usually interpreted as an indication that the model is misspecified, for instance by leaving out relevant levels, interactions among predictors, or in time series data by not allowing autocorrelation in the error structure.
Most programs rely on a Taylor expansion to linearize the model. The programs VARCL and MLn use a first-order Taylor expansion and marginal (quasi) likelihood (MQL1: P_{ij} predicted by fixed part only). MLn also offers the options of a second-order expansion and predictive or penalized (quasi) likelihood (PQL2: P_{ij} predicted by both fixed and random part).

Simulations by Rodriguez & Goldman (1995) show that marginal quasi likelihood with first-order Taylor expansion underestimates both the regression coefficients and the variance components, in some cases quite severely. Goldstein and Rasbash (1996) compare MQL1 and PQL2 by simulating data according to the worst performing dataset of Rodriguez and Goldman. In their simulation, the means of the MQL1 estimates for the fixed effects, from 200 simulation runs, were

underestimated by about 25%. The means of the MQL1 estimates for the random effects were underestimated by as much as 88%. Moreover, 54% of the level 2 variances were estimated as zero, while the population value is one. For the same 200 simulated datasets, the means of the PQL2 estimates for the fixed effects underestimated the population value by at most 3%, and for the random effects by at most 20%. None of the PQL2 variance estimates was estimated as zero. It appears that predictive quasi likelihood with second order Taylor expansion is sufficiently accurate for the regression coefficients, and in many cases good enough for the random parameters. However, with some data sets the PQL2 algorithm breaks down, and it is recommended to start with the simpler MQL1 approach to obtain good starting values for the more complicated PQL2 approach.
If anything, the analysis of proportions and binomial data requires larger samples that the analysis of normally distributed data. For proportions very close to 0 or 1 and small numbers of groups, Bayesian estimation using Gibbs sampling is coming into use; for a brief description see Goldstein (1995, p45).

5 Conclusion

As multilevel modeling becomes more widely used, it is important to gain an understanding of their limitations when the model assumptions are not fully met. This chapter, after establishing the need for multilevel modeling when data have a complex (hierarchical) error structure, focusses mostly on the effects of the sample sizes at different levels on accuracy and power of the statistical tests. A review of the available literature shows that estimates and tests for the regression coefficients are accurate with samples of modest sizes, but estimates and tests of the variances are not.

References

AFSHARTOUS, D. (1995): Determination of Sample Size for Multilevel Model Design. Paper, AERA Conference, San Francisco, 18-22 april 1995.

BARCIKOWSKI, R.S. (1981): Statistical Power with Group Mean as the Unit of Analysis. *Journal of Educational Statistics, 6, 267-285.*

BRYK, A.S. & RAUDENBUSH, S.W. (1992): Hierarchical Linear Models. Sage, Newbury Park, CA.

USING, F. (1993): Distribution Characteristics of Variance Estimates in Two-level Models. Department of Psychometrica and research Methodology, Leiden University, Leiden.

BURSTEIN, L. (1980): The Analysis of Multilevel Ddata in Education and Evaluation. *Review of Research in Education, 8, 158-233.*

DE LEEUW, J. & KREFT, ITA G.G. (1986): Random Coefficient Models for Multilevel Analysis. *Journal of Educational Statistics, 11, 57-85.*

FOTIU, R.P. (1989): A Comparison of the EM and Data Augmentation Algorithms on Simulated

Small Sample Hierarchical Data from Research on Education. Unpublished doctoral dissertation, Michigan State University, East Lansing.
GOLDSTEIN, H. (1995): Multilevel Statistical Models. Arnold, London.
GOLDSTEIN, H. & RASBASH, J. (1996): Improved Approximations for Multilevel Models with Binary Responses. Multilevel Models Project, University of London, London.
KIM, K.-S. (1990): Multilevel Data Analysis: a Comparison of Analytical Alternatives. Ph.D. Thesis, University of California, Los Angeles.
KISH, L. (1965): *Survey Sampling.* Wiley, New York.
KREFT, ITA G.G. (1996): Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies. California State University, Los Angeles.
LONGFORD, N. (1993): Random Coefficient Models. Clarendon Press, Oxford.
MOK, M. (1995): Sample Size requirements for 2-level Designs in Educational Research. Multilevel Models Project, University of London, London.
RAUDENBUSH, S.W. & BRYK, A.S. (1986): A Hierarchical Model for Studying School Effects. *Sociology of Education*, 59, 1-17.
RODRIGUEZ, G. & GOLDMAN, N. (1995): An Assessment of Estimation Procedures for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society, A-158, 73-90.*
SNIJDERS, T.A.B. & BOSKER, R. (1993): Modeled Variance in Two-level Models. *Journal of Educational Statistics, 18, 273-259.*
TATE, R. & WONGBUNDHIT, Y. (1983): Random versus Nonrandom Coefficient Models for Multilevel Analysis. *Journal of Educational Statistics, 8, 103-120.*
VAN DEN EEDEN, P. & HUETTNER, H.J.M. (1982): Multi-level Research. *Current Sociology, 30, 3, 1-117.*
VAN DER LEEDEN, R. & BUSING, F. (1994): First Iteration versus IGLS/RIGLS Estimates in Two-level Models: a Monte Carlo Study with ML3. Department of Psychometrica and research Methodology, Leiden University, Leiden.
VAN DER LEEDEN, R., BUSING, F., & MEIJER, E. (1997): Applications of Bootstrap Methods for Two-level Models. Paper, Multilevel Conference, Amsterdam, April 1-2, 1997.
WALD, A. (1943): Tests of Statistical Hypotheses Concerning several Parameters when the Number of Observations is Large. *Transactions of the American Mathematical Society, 54, 426-482.*