Sample Sizes for Multilevel Modeling

Cora J.M. Maas, Joop J. Hox

Corresponding author: C.J.M. Maas, E-mail: c.maas@fss.uu.nl

Address (both authors):

Department of Methodology and Statistics

Faculty of Social Sciences, Utrecht University

P.O.B. 80140

NL-3508 TC Utrecht, the Netherlands

Tel:/Fax: +31 30 2534594 / +31 30 253 5797

Abstract

The main problem of studying hierarchical systems, which often occur in social statistics, is the dependence of the observations at the lower levels. Multilevel analyzing programs account for this dependence and in recent years these programs have been widely accepted.

A problem in multilevel modeling is the question what constitutes a sufficient sample size for accurate estimation. In multilevel analysis, the restriction is often the higher-level sample size. In this paper, a simulation study is used to determine the influence of different sample sizes at the highest level on the accuracy of the estimates (regression coefficients and variances). In addition, the influence of other factors such as the lowest level sample size and different variance distributions between the levels (different intraclass correlations). The results show, that only a small sample size at level two (meaning a sample of 50 or less) leads to biased estimates of the second-level standard errors at the second level. In all of the other simulated conditions the estimates of both the regression coefficients, the variance components and the standard errors are unbiased and accurate.

Key words: Multilevel modeling, Sample size, Cluster sampling.

1. Introduction

Social research often involves problems that investigate the relationship between individual and society. The general concept is that individuals interact with the social contexts to which they belong, meaning that individual persons are influenced by the social groups or contexts to which they belong, and that the properties of those groups are in turn influenced by the individuals who make up that group. Generally, the individuals and the social groups are conceptualized as a hierarchical system of individuals and groups, with individuals and groups defined at separate levels of this hierarchical system.

Even if the analysis includes only variables at the lowest (individual) level, standard multivariate models are not appropriate. The hierarchical structure of the data creates problems, because the standard assumption of independent and identically distributed observations (i.i.d.) is generally not valid. Multilevel analysis techniques have been developed for the linear regression model (Bryk & Raudenbush, 1992; Goldstein, 1995), and specialized software is now widely available (Bryk, Raudenbush & Congdon, 1996; Rasbash & Woodhouse, 1995).

For example, assume that we have data from J groups, with a different number of respondents n_j in each group. On the respondent level, we have the outcome variable Y_{ij}. We have one explanatory variable X_{ij} on the respondent level, and one group level explanatory variable Z_j. To model these data, we have a separate regression model in each group as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} . \qquad (1)$$

The variation of the regression coefficients β_j is modeled by a group level regression model, as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}\,, \tag{2}$$

and

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

The individual-level residuals e_{ij} are assumed to have a normal distribution with mean zero and variance σ_e^2. The group-level residuals u_{0j} and u_{1j} are assumed to have a multivariate normal distribution with expectation zero, and to be independent from the residual errors e_{ij}. The variance of the residual errors u_{0j} is specified as σ_{00}, and the variance of the residual errors u_{1j} is specified as σ_{11}.

This model can be written as one single regression model by substituting equations (2) and (3) into equation (1). Substitution and earranging terms gives:

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

The segment [$\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij}$] in equation (4) contains all the fixed coefficients; it is the fixed (or deterministic) part of the model. The segment [$u_{0j} + u_{1j} X_{ij} + e_{ij}$] in equation (4) contains all the random error terms; it is the random (or stochastic) part of the model. The term

Z_jX_{ij} is an interaction term that appears in the model because of modeling the varying regression slope β_{1j} of respondent level variable X_{ij} with the group level variable Z_j.

Multilevel models are needed because grouped data violate the assumption of independence of all observations. The amount of dependence can be expressed as the intraclass correlation ρ. In the multilevel model, the intraclass correlation is estimated by specifying an empty model, as follows:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \quad (5)$$

Using this model we can estimate the intraclass correlation ρ by the equation

$$\rho = \sigma_{00} / (\sigma_{00} + \sigma_e^2). \quad (6)$$

The maximum likelihood estimation methods commonly used in multilevel analysis are asymptotic, which translates to the assumption that the sample size is large. This arouses questions about the accuracy of the various estimation methods with relatively small sample sizes. The question is of interest for survey researchers, because multilevel models may be useful in their own right, e.g., as in their use to model interview or region effects (cf. O'Muircheartaigh & Campanelli, 1999; Pickery & Loosvelt, 1998), but also because multilevel techniques offer one way of dealing with clustered or stratified data (Goldstein & Silver, 1989; Snijders, 2001). A recent simulation study on multilevel structural equation modeling (Hox & Maas, 2001) suggests that the size of the intraclass correlation (ICC) also affects the accuracy of the estimates. Therefore, in our simulation, we have varied not only the sample size at the individual

and the group level, but also the intraclass correlation. In general, what is at issue in multilevel modeling is not so much the intraclass correlation, but the *design effect*, which indicates how much the standard errors are underestimated (Kish, 1965). In cluster samples, the design effect is approximately equal to 1+(average cluster size-1)*ICC. If the design effect is smaller than two, using single level analysis on multilevel data does not seem to lead to overly misleading results (Muthén & Satorra, 1995). In our simulation setup, we have chosen values for the ICC and group sizes that make the design effect larger than two in all simulated conditions.

2. Review of existing research

There are some simulation studies on this topic, which mostly investigate the accuracy of the fixed and random parameters with small sample sizes at either the individual or the group level. Comparatively less research investigates the accuracy of the standard errors used to test specific model parameters.

2.1 Accuracy of fixed parameters and their standard errors

The estimates for the regression coefficients appear generally unbiased, for Ordinary Least Squares (OLS), Generalized Least Squares (GLS), as well as Maximum Likelihood estimation (Van der Leeden & Busing, 1994; Van der Leeden et al., 1997). OLS estimates are less efficient; they have a larger sampling error. Kreft (1996), reanalyzing results from Kim (1990), estimates that OLS estimates are about 90% efficient.

The OLS based standard errors are severely biased downwards (Snijders & Bosker, 1999). The asymptotic Wald tests, used in most multilevel software, assume large samples. 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 standard errors have a small downward bias. GLS estimates of fixed parameters and their standard errors are somewhat less accurate than ML estimates, but workable. In general, a large number of groups appears more important than a large number of individuals per group.

2.2 Accuracy of random parameters and their standard errors

Estimates of the residual error at the lowest level are generally very accurate. The group level variance components are sometimes underestimated. Simulation studies by Busing (1993) and Van der Leeden and Busing (1994) show that GLS variance estimates are less accurate than ML estimates. The same simulations also indicate that for accurate group level variance estimates many groups (more than 100) are needed (cf. Afshartous, 1995). In contrast, Browne and Draper (2000) show that with as few as six to twelve groups, Restricted ML (RML) estimation provides good variance estimates, and with as few as 48 groups, Full ML (FML) estimation also produces good variance estimates. We will come back to these apparently contradictory results in our discussion.

The simulations by Van der Leeden et al. (1997) show that the standard errors used for the Wald test of the variance components are generally estimated too small, with RML again more accurate than FML. Symmetric confidence intervals around the estimated value also do not perform well. Browne and Draper (2000) report similar results. Typically, with 24-30 groups,

Browne and Draper report an operating alpha level of about 9%, and with 48-50 groups about 8%. Again, In general, a large number of groups appears more important than a large number of individuals per group.

3. Method

3.1 The simulation model and procedure

We use a simple two-level model, with one explanatory variable at the individual level and one explanatory variable at the group level, conforming to equation (4), which is repeated here:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{1j} X_{ij} + u_{0j} + e_{ij} . \qquad \text{(4, repeated)}$$

Three conditions are varied in the simulation: (1) Number of Groups (NG: three conditions, NG=30,50,100), (2) Group Size (GS: three conditions, GS=5, 30, 50), and (3) Intraclass Correlation (ICC: three conditions, ICC=0.1, 0.2, 0.3).[1]

The number of groups is chosen so that the highest number should be sufficient given the simulations by Van der Leeden et al. (1997). In practice, 50 groups is a frequently occurring number in organizational and school research, and 30 is the smallest number according to Kreft and De Leeuw (Kreft & De Leeuw, 1998). Similarly, the group sizes are chosen so that the highest number should be sufficient. A group size of 30 is normal in educational research, and a group size of five is normal in family research and in longitudinal research, where the measurement occasions form the lowest level. The Intra Class Correlations (ICC's) span the

customary level of intraclass correlation coefficients found in studies where the groups are formed by households (Gulliford, Ukoumunne & Chinn, 1999).

There are 3×3×3=27 conditions. For each condition, we generated 1000 simulated data sets, assuming normally distributed residuals. The multilevel regression model, like its single-level counterpart, assumes that the explanatory variables are fixed. Therefore, a set of X and Z values are generated from a standard normal distribution to fulfill the requirements of the simulation condition with the smallest total sample size. In the condition with the larger sample sizes, these values are repeated. This ensures that in all simulated conditions the joint distribution of X and Z are the same. The regression coefficients are specified as follows: 1.00 for the intercept, and 0.3 (a medium effect size, cf. Cohen, 1988) for all regression slopes. The residual variance σ_e^2 at the lowest level is 0.5. The residual variance σ_{00} follows from the specification of the ICC and σ_e^2, given formula (6). Busing (1993) shows that the effects for the intercept variance σ_{00} and the slope variance σ_{11} are similar; hence, we chose to use the value of σ_{00} also for σ_{11}. To simplify the simulation model, without loss of generality, the covariance between the two *u*-terms is assumed equal to zero.

Two Maximum Likelihood functions are common in multilevel estimation: Full ML (FML) and Restricted ML (RML). We use RML, since this is always at least as good as FML, and sometimes better, especially in estimating variance components (Browne, 1998). The software MLwiN (Rasbash et al., 2000) was used for both simulation and estimation.

3.2 Variables and analysis

The percentage relative bias is used to indicate the accuracy of the parameter estimates (factor loadings and residual variances). Let $\hat{\theta}$ be the estimate of the population parameter θ, then the percentage relative bias is given by $\left(\left(\hat{\theta}-\theta\right)/\theta\right)\times 100\%$. The accuracy of the standard errors is investigated by analyzing the observed coverage of the 95% confidence interval. Since the total sample size for each analysis is 27000 simulated conditions, the power is huge. As a result, at the standard significance level of alpha=0.05, extremely small effects become significant. Therefore, our criterion for significance is an alpha = 0.01 for the main effects of the simulated conditions. The interactions are tested blockwise (2-way, 3-way), with a Bonferroni correction added for separate interaction effects. Even at this stricter level of significance, some of the statistically significant biases correspond to differences in parameter estimates that do not show up before the third decimal place. These small effects are discussed in the text, but not included in the various tables.

4. Results

4.1 Convergence and inadmissible solutions

The estimation procedure converged in all 27000 simulated data sets. The estimation procedure in MLwiN can and sometimes does lead to negative variance estimates. Such solutions are inadmissible, and common procedure is to constrain such estimates to the boundary value of zero. However, all 27000 simulated data sets produced only admissible solutions.

4.2 Parameter estimates

The fixed parameter estimates, the intercept and regression slopes, have a negligible bias. The average bias is smaller than 0.05%. The largest bias was found in the condition with the smallest sample sizes in combination with the highest ICC: there the percentage relative bias was 0.3%. This is of course extremely small. Moreover, there are no statistically significant differences in bias across the simulated conditions.

The estimates of the random parameters, the variance components, also have a negligible bias. The average bias is smaller than 0.05%. The largest bias was found in the condition with the smallest sample sizes in combination with the highest ICC: there the percentage relative bias was 0.3%.

4.3 Standard errors

To assess the accuracy of the standard errors, for each parameter in each simulated data set the 95% confidence interval was established using the asymptotic standard normal distribution (cf. Goldstein, 1995; Longford, 1993). For each parameter a non-coverage indicator variable was set up which is equal to zero if its true value is in the confidence interval, and equal to one if its true value is outside the confidence interval. The effect of the different simulated conditions on the non-coverage was analyzed using logistic regression on these indicator variables.

The non-coverage of both fixed and random effects is significantly affected by the number of groups and by the group size. Non-coverage is not sensitive to the Intraclass correlation. The

effect of the number of groups on the non-coverage is presented in Table 1, and the effect of the group size on non-coverage is presented in Table 2.

==========

Tables 1 and 2 about here

==========

Table 1 shows that the effect of the number of groups on the standard errors of the fixed regression coefficients is small. With 30 groups, the non-coverage rate is 6.0% for the regression coefficient and 6.4% for the intercept, while the nominal non-coverage rate is 5%. We regard this difference as trivial. The effect of the number of groups on the standard errors of the random variance components is definitely larger. With 30 groups, the non-coverage rate for the second-level intercept variance is 8.9%, and the non-coverage rate for the second-level slope variance is 8.8%. Although the coverage is not grotesquely wrong, the 95% confidence interval is clearly too short. The amount of non-coverage here implies that the standard errors for the second-level variance components are estimated about 15% too small. With 50 groups, the effects are smaller, but still not negligible. The non-coverage rates of 7.4% and 7.2% imply that the standard errors are estimated 9% too small.

Table 2 shows that the non-coverage of the lowest level variance is improved when the group size increases. The non-coverage of the second-level variances does not improve when the group size increases.

5 Summary and discussion

Concluding, the point estimates of both the fixed regression coefficients and the random variance components are all estimated without bias, in all of the simulated conditions. The standard errors of the fixed regression coefficients are also estimated accurately. Only the standard errors of the second-level variances are estimated too small when the number of groups is substantially lower than 100. With 30 groups, the standard errors are estimated about 15% too small, resulting in a non-coverage rate of almost 9%, instead of 5%. With 50 groups, the non-coverage drops to about 7.3%. This is clearly different from the nominal 5%, but in practice acceptable.

Our simulation results indicate that Maximum Likelihood estimation for multilevel models leads to unbiased point estimates. The asymptotic standard errors are also accurate, with the exception of the standard errors for second-level variances in the case of a small sample of groups (less than 50). These results differ to some extent from the simulation results reported by Busing (1993) and Van der Leeden and Busing (1994). They conclude that for small sample sizes the standard errors and corresponding statistical tests are badly biased, so that about100 groups are needed for accurate estimation of variance components. According to our simulations about fifty groups appears sufficient. However, they use a different simulation design. Busing (1993) uses much higher intraclass correlations, up to 0.80, which are unlikely to occur in actual data. In addition, the simulated second-level sample sizes are much smaller, starting at a sample of 10 groups with 5 observations each. For these simulated conditions, they report both convergence problems and biased standard statistical tests, especially for the variance components. However, their simulation results are comparable when only those conditions are considered that are similar to the conditions in our simulations and those of Browne and Draper (2000).

To investigate the limits of our results, we carried out two additional small simulation studies. First, we increased the population values of the residual variances, which decreases the amount of explained variance in the population model. The results were very close to the results reported above. Second, we carried out a simulation with only ten groups of group size five. This simulation was inspired by the dissimilar results of Busing (1993), and by a statement in Snijders and Bosker (1999, p44) that multilevel modeling becomes attractive when the number of groups is larger than ten. This is a very small second-level sample size, but given our simulation results not impossibly small. In this simulation the fixed regression coefficients and variance components were still estimated without bias, except for the second-level variance components when the ICC was low (0.10): there the bias is 25% upwards. The standard errors are now all estimated too small. The non-coverage rates for the fixed effects in this case range between 5.7% and 9.7%, and for the second-level variances they range between 16.3% and 30.4%. Although the standard errors of the fixed effects are still reasonable, the standard errors of the second-level variances are clearly unacceptable. It would seem that having as few as ten groups is too few. If one is interested only in the fixed regression coefficients, it appears feasible, but we would advise to use bootstrapping or other simulation-based methods to assess the sampling variability.

6 References

Afshartous, D. (1995). *Determination of sample size for multilevel model design.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Browne, W.J. (1998). Applying MCMC methods to multilevel models. Unpublished Ph.D. Thesis, University of Bath, UK.

Browne, W.J. & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computaional Statistics*, 15, 391-420.

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

Bryk, A.S., Raudenbush, S.W. & Congdon, R.T. (1996). *HLM. Hierarchical linear and nonlinear modeling with the HLM/2L and HLM?3L programs.* Chicago: Scientific Software International.

Busing, F. (1993). Distribution characteristics of variance estimates in two-level models. Unpublished manuscript. Leiden: Department of Psychometrics and Research Methodology, Leiden University.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Mahwah, NJ: Erlbaum.

O'Muircheartaigh, C. & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, 162, 437-446.

Goldstein, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold/New York: Halsted.

Goldstein, H. & Silver, R. (1989) Multilevel and Multivariate Models in Survey Analysis. In Analysis of Complex Surveys, eds. C.J. Skinner, D. Holt & T.M. Smith. New York: Wiley.

Gulliford, M.C., Ukoumunne, O.C. & Chinn, S. (1999). Components of Variance and Intraclass Correlations for the Design of Community-based Surveys and Intervention Studies.

American Journal of Epidemiology, 149, 876-883. Longford, N.T. (1993). *Random coefficient models*. Oxford: Clarendon Press.

Hox, J.J. & Maas, C.J.M. (2001). The Accuracy of Multilevel Structural Equation Modeling With Pseudobalanced Groups and Small Samples. *Structural Equation Modeling,* 8, 157-174.

Kim K.-S. (1990). Multilevel data analysis: A comparison of analytical alternatives. Unpublished Ph.D. thesis, University of California, Los Angeles.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kreft, I.G.G. (1996): Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies. Unpublished Report, California State University, Los Angeles. Available at http://ioe.ac.uk/multilevel.

Kreft & De Leeuw (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Longford, N.T. (1993). *Random coefficient models*. Oxford: Clarendon Press.

Muthén, B. & Satorra, A. (1995). Complex sample data in structural equation modeling. In P.V. Marsden (Ed.), *Sociological methodology* (pp. 267-316). Oxford, England: Blackwell.

O'Muircheartaigh, C. & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, 162, 437-446.

Pickery, J. & Loosveldt, G. (1998). The impact of respondent and interviewer characteristics on the number of 'no opinion'answers. *Quality & Quantity*, 32, 31-45.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. & Lewis, T. (2000). A user's guide to MlwiN. London: Multilevel Models Project, University of London.

Rasbash, J., & Woodhouse, G. (1995). *MLn command reference.* London: Institute of Education, University of London.

Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* Sage Publications: London-Thousand Oaks-New Delhi.

Snijders, T.A.B. (2001). Sampling. In A. Leyland & H. Goldstein (eds.). *Multilevel Modelling of Health Statistics*. New York: Wiley.

Van der Leeden, R., & Busing, F. (1994). First iteration versus IGLS?RIGLS estimates in two-level models: A Monte Carlo study with ML3. Unpublished manuscript, Department of Psychometrics and Research Methodology, Leiden University.

Van der Leeden, R., Busing, F., & Meijer, E. (1997): Applications of Bootstrap Methods for Two-level Models. Unpublished paper, Multilevel Conference, Amsterdam, April 1-2, 1997.

Table 1. Influence of the Number of Groups on the non-coverage of the 95% confidence interval

Parameter	Number of Groups			
	30	50	100	p-value
U0	0.089	0.074	0.060	.0000
U1	0.088	0.072	0.057	.0000
E0	0.058	0.056	0.049	.0102
INT	0.064	0.057	0.053	.0057
X	0.060	0.057	0.050	.0058

Table 2 Influence of the Group Size on the non-coverage of the 95% confidence interval

Parameter	Group Size			
	5	30	50	*p*-value
U0	0.074	0.075	0.074	.9419
U1	0.078	0.066	0.072	.0080
E0	0.061	0.051	0.051	.0055

Footnote:

[1] Corrections for clustering based on the design effect (Kish, 1965) assume equal group sizes; multilevel analysis does not. We carried out some preliminary simulations to assess if having balanced or unbalanced groups has any influence on multilevel ML estimates. There was no effect of balance on the multilevel estimates or their standard errors.