Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 46 (2004) 427–440
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

The influence of violations of assumptions on multilevel parameter estimates and their standard errors

Cora J.M. Maas , Joop J. Hox*

Department of Methodology and Statistics, Utrecht University, Netherlands

Received 5 August 2003; received in revised form 5 August 2003

Abstract

A crucial problem in the statistical analysis of hierarchically structured data is the dependence of the observations at the lower levels. Multilevel modeling programs account for this dependence and in recent years these programs have been widely accepted. One of the major assumptions of the tests of significance used in the multilevel programs is normality of the error distributions involved. Simulations were used to assess how important this assumption is for the accuracy of multilevel parameter estimates and their standard errors. Simulations varied the number of groups, the group size, and the intraclass correlation, with the second level residual errors following one of three non-normal distributions. In addition asymptotic maximum likelihood standard errors are compared to robust (Huber/White) standard errors.

The results show that non-normal residuals at the second level of the model have little or no effect on the parameter estimates. For the fixed parameters, both the maximum likelihood-based standard errors and the robust standard errors are accurate. For the parameters in the random part of the model, the maximum likelihood-based standard errors at the lowest level are accurate, while the robust standard errors are often overcorrected. The standard errors of the variances of the level-two random effects are highly inaccurate, although the robust errors do perform better than the maximum likelihood errors. For good accuracy, robust standard errors need at least 100 groups. Thus, using robust standard errors as a diagnostic tool seems to be preferable to simply relying on them to solve the problem.

Keywords: Multilevel modeling; Maximum likelihood; (robust) Standard errors; Sandwich estimate; Huber/White correction

* Corresponding author. Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University. P.O.B. 80140, NL-3508 TC Utrecht, Netherlands. Tel.: +31-30-253-9236; fax: +31-30-253-5797.
E-mail addresses: c.maas@fss.uu.nl (C.J.M. Maas), j.hox@fss.uu.nl (J.J. Hox).

0167-9473/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2003.08.006

1. Introduction

Social research often involves problems that investigate the relationship between individual and society. The general concept is that individuals interact with their social contexts, meaning that individual persons are influenced by the social groups or contexts, 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, with individuals and groups defined at separate levels of this hierarchical system.

Standard multivariate models are not appropriate for the analysis of such hierarchical systems, even if the analysis includes only variables at the lowest (individual) level, because the standard assumption of independent and identically distributed observations is generally not valid. The consequences of using uni-level analysis methods on multilevel data are well known: the parameter estimates are unbiased but inefficient, and the standard errors are negatively biased, which results in spuriously 'significant' effects (cf. de Leeuw and Kreft, 1986; Snijders and Bosker, 1999; Hox, 1998, 2002). Multilevel analysis techniques have been developed for the linear regression model (Bryk and Raudenbush, 1992; Goldstein, 1995), and specialized software is now widely available (Raudenbush et al., 2000; Rasbash et al., 2000).

The assumptions underlying the multilevel regression model are similar to the assumptions in ordinary multiple regression analysis: linear relationships, homoscedasticity, and normal distribution of the residuals. In ordinary multiple regression, it is known that moderate violations of these assumptions do not lead to highly inaccurate parameter estimates or standard errors. Thus, provided that the sample size is not too small, standard multiple regression analysis can be regarded as a robust analysis method (cf. Tabachnick and Fidell, 1996). In the case of severe violations, a variety of statistical methods for correcting heteroscedasticity are available (Scott Long and Ervin, 2000). Multilevel regression analysis has the advantage that heteroscedasticity can also be modeled directly (cf. Goldstein, 1995, pp. 48–57).

The maximum likelihood estimation methods used commonly in multilevel analysis are asymptotic, which translates to the assumption that the sample size is large. This raises questions about the accuracy of the various estimation methods with relatively small sample sizes. This concerns especially the higher level(s), because the sample size at the highest level (the sample of groups) is always smaller than the sample size at the lowest level. A large simulation by Maas and Hox (2003) finds that the standard errors for the regression coefficients are slightly biased downwards if the number of groups is less than 50. With 30 groups, they report an operative alpha level of 6.4% while the nominal significance level is 5%. Similarly, simulations by Van der Leeden and 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.

Sometimes it is possible to obtain more nearly normal distributions by transforming the outcome variable. If this is undesirable or even impossible, another method to obtain better tests and confidence intervals is to correct the asymptotic standard errors. One correction method to produce robust standard errors is the so-called Huber/White

or sandwich estimator (Huber, 1967; White, 1982), which is available in several of the available multilevel analysis programs (e.g., Raudenbush et al., 2000; Rasbash et al., 2000).

In this paper we look more precisely at the consequences of the violation of the assumption of normally distributed errors at the second level of the multilevel regression model. Specifically, we use simulation to answer the following two questions: (1) what group level sample size can be considered adequate for reliable assessment of sampling variability when the assumption of normally distributed residuals is not met, and (2) how well do the asymptotic and the sandwich estimators perform when the assumption of normally distributed residuals is not met.

2. The multilevel regression model

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:

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

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

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

This model can be written as a single regression model by substituting Eqs. (2) and (3) into Eq. (1). Substitution and rearranging 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 $(\gamma_{00}+\gamma_{10}X_{ij}+\gamma_{01}Z_j+\gamma_{11}Z_jX_{ij})$ in Eq. (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 Eq. (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 a null model, as follows:

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

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

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

where σ_e^2 is the variance of the individual-level residuals and σ_{00} the variance of the residual errors u_{0j}.

The assumptions underlying the multilevel model are linear relations, a normal distribution for the individual-level residuals e_{ij} (with mean zero and variance σ_e^2), and a multivariate normal distribution for the group-level residuals u_{0j} and u_{1j} (with expectation zero and variances σ_{00} and σ_{11}; these residuals are assumed independent from the residual errors e_{ij}).

3. Maximum likelihood estimation

The usual estimation method for the multilevel regression model is maximum likelihood (ML) estimation (cf. Eliason, 1993). One important assumption underlying this estimation method is normality of the error distributions. When the residual errors are not normally distributed, the parameter estimates produced by the ML method are still consistent and asymptotically unbiased. However, the asymptotic standard errors are incorrect. Significance tests and confidence intervals can thus not be trusted (Goldstein, 1995). This problem does not completely vanish when the sample gets larger.

3.1. The sandwich estimator

One method to obtain better tests and confidence intervals is to correct the asymptotic standard errors of the fixed and random parameters, using the so-called Huber/White or sandwich estimator (Huber, 1967; White, 1982). In the ML approach, the usual estimator of the sampling variances and covariances is the inverse of the information matrix (Hessian matrix, cf. Eliason, 1993). Using matrix notation, the asymptotic variance–covariance matrix of the estimated regression coefficients can be written as follows:

$$\mathbf{V}_A(\hat{\beta}) = \mathbf{H}^{-1}, \tag{7}$$

where $\mathbf{V}_A$ is the asymptotic covariance matrix of the regression coefficients, and $\mathbf{H}$ is the Hessian matrix. The Huber/White estimator is given as

$$\mathbf{V}_R(\hat{\beta}) = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-1}, \tag{8}$$

where $\mathbf{V}_R$ is the robust covariance matrix of the regression coefficients, and $\mathbf{C}$ is a correction matrix. The correction matrix, which is 'sandwiched' between the two $\mathbf{H}^{-1}$ terms, is based on the observed raw residuals. Details of the Huber/White correction for the multilevel model are given by Goldstein (1995) and Raudenbush and Bryk (2002). If the residuals follow a normal distribution, $\mathbf{V}_A$ and $\mathbf{V}_R$ are both consistent estimators of the covariances of the regression coefficients, but the model-based asymptotic covariance matrix $\mathbf{V}_A$ is more efficient because it leads to the smallest standard errors. However, when the residuals do not follow a normal distribution, the model-based

asymptotic covariance matrix is not correct, while the observed residuals-based sandwich estimator $\mathbf{V}_R$ is still a consistent estimator of the covariances of the regression coefficients. This makes inference based on the robust standard errors less dependent on the assumption of normality, at the cost of sacrificing some statistical power and possibly the good approximation of the nominal significance level.

3.2. Some well-known factors influencing parameter estimates and standard errors

Since the ML estimation methods are asymptotic, the assumption is that the sample size is large. With small sample sizes, the estimates for the fixed regression coefficients appear generally unbiased (Maas and Hox, 2003). When assumptions of normality and large samples are not met, the standard errors of the fixed parameters have a small downward bias (Van der Leeden and Busing, 1994; Van der Leeden et al., 1997). 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) indicate that when high accuracy is wanted for the group level variance estimates, many groups (more than 100) are needed (cf. Afshartous, 1995). In contrast, Browne and Draper (2000) show that in some cases with as few as 6–12 groups, restricted ML (RML) estimation can provide useful variance estimates, and with as few as 48 groups, full ML (FML) estimation also produces useful variance estimates. Our own simulations (Maas and Hox, 2003) with normal data and using RML indicate that about 50 groups are needed to have both good variance estimates for the parameters in the random part of the model, and accurate standard errors for these variance estimates.

A simulation study of Maas and Hox (2003) shows that only a small sample size at the group level (meaning a sample of 50 or less) leads to biased estimates of the group-level standard errors. Furthermore, the simulations by Van der Leeden et al. (1997) show that the standard errors 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%. A large number of groups is more important than a large number of individuals per group.

A recent simulation study on multilevel structural equation modeling (Hox and 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 ICC. In general, what is at issue in multilevel modeling is not so much the ICC, 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 and Satorra, 1995). We have chosen values for the ICC and group sizes that make the design effect larger than two in all simulated conditions.

4. Method

4.1. 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 Eq. (4), which is repeated here

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

Four conditions are varied in the simulation: (1) Number of groups (NG: three conditions, NG = 30, 50 and 100), (2) group size (GS: three conditions, GS = 5, 30 and 50), (3) intraclass Correlation (ICC: three conditions, ICC = 0.1, 0.2 and 0.3; note that the ICC varies with the X, so these ICCs apply to the average case where $X = 0$) and type of level-2 residual distribution (three conditions, described below).

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 of groups according to Kreft and 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 ICCs span the customary level of ICC coefficients found in studies where the groups are formed by households (Gulliford et al., 1999).

There are $3 \times 3 \times 3 \times 3 = 81$ 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 conditions 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 Eq. (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, the covariance between the two u-terms is assumed equal to zero. Given the parameter values, the simulation procedure generates the residual errors e_{ij} and $u_{.j}$. To investigate the influence of non-normally distributed errors we replaced the second-level residuals with residuals generated from a non-normal distribution, transformed to have a mean of zero and a standard deviation corresponding to the correct population value. The three non-normal distributions used were a chi-square distribution with one degree of freedom, which is markedly skewed, a uniform distribution, which has heavy tails compared to the normal distribution, and a Laplace distribution with location parameter zero and scale parameter one, which is symmetric with smaller tails than a normal distribution (Evans et al., 1993). We

consider each of these distributions a different but large deviation of the assumption of having a multivariate normal distribution for the second-level residuals.

Two ML functions are common in multilevel estimation: FML and 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 analyses are carried out twice, once with asymptotic ML-based standard errors, and once with robust Huber/White standard errors. The software MLwiN (Rasbash et al., 2000) was used for both simulation and estimation. In this program the correction of the sandwich estimation is based on the cross-product matrix of the residuals, taking the multilevel structure of the data into account.

4.2. Variables and analysis

To indicate the accuracy of the parameter estimates (regression coefficients and residual variances) the percentage relative bias is used. Let $\hat{\theta}$ be the estimate of the population parameter θ, then the percentage relative bias is given by $100 \times \hat{\theta}/\theta$. 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 27,000 simulated conditions, the power is huge. As a result, at the standard significance level of $\alpha = 0.05$, extremely small effects become significant. Therefore, we set our criterion for significance to $\alpha = 0.001$ for the main effects of the simulated conditions. To compare different conditions we used Anova.

5. Results

5.1. Convergence and inadmissible solutions

The estimation procedure converged in all $3 \times 27,000 = 81,000$ 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 simulated data sets produced only admissible solutions.

5.2. Percentage relative bias

For across all 27 conditions the mean relative bias is calculated. Tested is whether this relative bias differs from one, with an α of 0.001. The p-values in the table are Bonferroni-corrected (the 2-tailed p-value is multiplied by 7, because 7 mean parameters are tested). The percentage relative bias is the same for the ML- and the robust estimations, because we investigate the parameter estimates and not their standard errors. There was only one significant effect of the lower level variance. This was for the chi-squared residual errors in the "worst" condition, meaning 30 groups with five individuals and an ICC of 0.1. This significant effect is totally irrelevant (variance

Table 1
Relative bias of the parameter estimates chi-squared residuals[a] ($\alpha = 0.001$)

	Relative bias	Population value	Estimate	p-value
Intercept	1.002	1.00	1.002	1.000
X	0.990	0.30	0.297	1.000
Z	0.997	0.30	0.299	1.000
XZ	1.002	0.30	0.301	1.000
E_0	0.984	0.50	0.492	0.001*
U_0	1.116	0.056	0.063	0.005
U_1	1.035	0.056	0.058	1.000

[a]Uniform and Laplace residuals: no difference from population value.
*sign.

Table 2
Coverage of the 95% confidence interval for the main fixed effects ($0.9260 < \text{CI} < 0.9740; \alpha = 0.001$)

	ML-estimation	Robust estimation
Intercept	0.9322/0.9420/0.9454	0.9291/0.9388/0.9430
X	0.9262/0.9461/0.9453	0.9229*/0.9432/0.9419
Z	0.9458/0.9454/0.9509	0.9402/0.9364/0.9415
XZ	0.9484/0.9521/0.9491	0.9365/0.9409/0.9382

First: Chi^2, second: uniform, third: Laplace.
*sign.

estimated as 0.492 instead of 0.50). The results of this "worst" condition are given in Table 1. All other parameter estimates in all conditions were estimated without bias.

5.3. Confidence intervals

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). The coverage of both fixed and random parameters is significantly affected by the number of groups and by the group size, coverage of the random parameters also by the ICC.

The coverage of the 95% confidence interval for the main fixed effects for all simulated conditions is presented in Table 2. In 1000 simulated data sets, for $\alpha = 0.001$ the confidence interval of the estimated confidence interval (CI) equals: $0.9260 < \text{CI} < 0.9740$. Values of the coverage outside this interval indicate a significant deviation from the statistical norm. Only in the case of the robust estimation with chi-squared residuals there is one small significant effect.

In Table 3 the coverage in the three conditions for the fixed effects are compared using the nonparametric Kruskal–Wallis Test. There are effects of the number of groups and of the group size. With respect to the group size, the results are as expected: larger group sizes lead to a closer approximation of the nominal coverage. The number of

Table 3
Significance of the effect on coverage of the 95% confidence interval for the three conditions for the fixed effects (first the p-value for the ML-estimation; second for the robust estimation)

	Intercept	*X*	*Z*	*XZ*
Number groups				
Chi^2	0.0144/0.0004*	0.0000*/0.0000*	0.0240/0.0000*	0.0964/0.5880
Uniform	0.0532/0.0236	0.0468/0.0064	1.000/0.2052	1.000/0.3224
Laplace	0.0512/0.0052	0.0940/0.0064	1.000/0.0172	1.000/0.0004*
Group size				
Chi^2	0.0000*/0.0000*	0.0000*/0.0000*	0.0004*/0.1704	0.0040*/0.0016
Uniform	0.6612/0.2600	0.0020/0.0008*	0.2180/0.8072	0.0000*/0.0020
Laplace	0.9992/1.000	0.0252/0076	0.0036/0.0544	0.0508/0.0000*
ICC				
Chi^2	0.3636/0.3344	0.3748/0.3880	1.000/1.000	1.000/1.000
Uniform	1.000/1.000	1.000/1.000	1.000/1.000	1.000/1.000
Laplace	1.000/1.000	1.000/1.000	1.000/1.000	1.000/1.000

*sign.

groups has more effect on the coverage bias when the robust standard errors are used than the ML-standard errors, both with the chi-squared and the Laplace residuals.

The effect of the number of groups and of the group size on the coverage is presented in Table 4. The coverage intervals reported in Table 4 are significantly different from the nominal coverage if they lie outside the interval $0.9260 < \mathrm{CI} < 0.9740$. The significant effects are relatively small, and mostly due to number of groups with the chi-squared residuals. In the second part of Table 4 we see that when the group sizes become larger, the *p*-values of the lower level regression coefficients become significant. This seems anomalous, but it is the predictable effect of a larger design effect resulting from the combination of a specific ICC value with a larger group size.

The coverage of the 95% confidence interval for the variance estimates is presented in Table 5 ($0.9265 < \mathrm{CI} < 0.9735$). The ML-estimations give correct estimates for the lowest level parameter. At the second level we observe large deviations from the nominal coverage (coverage of 0.66 and 0.64). The robust estimation produces overcorrected standard errors at the lowest level (coverage of 0.99 instead of 0.95) and still large deviations at the second level (coverage of 0.87 and 0.85). However, these deviations are considerably smaller than the deviations of the ML standard errors. We find the largest deviations with the chi-squared residuals, but there is still considerable bias with the Laplace-distributed residuals.

In Table 6 the effects of the three conditions on the coverage for the variance estimates are compared using the nonparametric Kruskal–Wallis Test. All three conditions have significant effects, mostly for the chi-squared residuals, but also for uniform residuals.

The effects of the number of groups on the coverage are presented in the first part of Table 7; the effect of the group size on coverage is presented in the second part and the effect of the ICC in the third part. The coverage intervals reported in Table 7 are significantly different from the nominal coverage if they lie outside the interval

Table 4
Effect of number of groups and group size on coverage of the 95% confidence interval for on the fixed effects (first the p-value for the ML-estimation; second for the robust estimation)

	Intercept	X	Z	XZ
NG				
30				
Chi^2	0.9271/0.9214*	0.9171*/0.9120*	0.9397/0.9306	0.9529/0.9367
Uniform	0.9417/0.9370	0.9404/0.9361	0.9481/0.9317	0.9534/0.9370
Laplace	0.9436/0.9397	0.9403/0.9351	0.9511/0.9352	0.9494/0.9323
50				
Chi^2	0.9302/0.9279	0.9246*/0.9214*	0.9498/0.9439	0.9439/0.9329
Uniform	0.9371/0.9342	0.9499/0.9474	0.9440/0.9371	0.9528/0.9409
Laplace	0.9417/0.9390	0.9459/0.9430	0.9523/0.9428	0.9480/0.9351
100				
Chi^2	0.9392/0.9379	0.9370/0.9353	0.9480/0.9462	0.9484/0.9400
Uniform	0.9473/0.9452	0.9481/0.9461	0.9441/0.9404	0.9502/0.9449
Laplace	0.9511/0.9502	0.9496/0.9474	0.9492/0.9466	0.9499/0.9472
GS				
5				
Chi^2	0.9422/0.9390	0.9378/0.9328	0.9543/0.9440	0.9413/0.9288
Uniform	0.9390/0.9353	0.9403/0.9362	0.9501/0.9386	0.9457/0.9352
Laplace	0.9487/0.9453	0.9400/0.9359	0.9576/0.9469	0.9436/0.9260*
30				
Chi^2	0.9266/0.9236	0.9247*/0.9221*	0.9414/0.9353	0.9519/0.9431
Uniform	0.9416/0.9377	0.9532/0.9506	0.9428/0.9380	0.9624/0.9486
Laplace	0.9442/0.9426	0.9450/0.9414	0.9458/0.9367	0.9524/0.9440
50				
Chi^2	0.9278/0.9247*	0.9162*/0.9139*	0.9417/0.9413	0.9520/0.9377
Uniform	0.9456/0.9434	0.9449/0.9429	0.9433/0.9327	0.9483/0.9390
Laplace	0.9434/0.9410	0.9508/0.9482	0.9493/0.9410	0.9513/0.9447

*sign.

Table 5
Coverage of the 95% confidence interval for the main random effects ($0.9273 < \mathrm{CI} < 0.9727; \alpha = 0.001$)

	ML-estimation	Robust estimation
E_0	0.9520/0.9503/0.9501	0.9901*/0.9884*/0.9889*
U_0	0.6632*/0.9663/0.8381*	0.8693*/0.9763*/0.9329
U_1	0.6427*/0.9661/0.8253*	0.8524*/0.9776*/0.9248*

First: Chi^2, second: uniform, third: Laplace.
*sign.

$0.9265 < \mathrm{CI} < 0.9735$. In all simulated conditions the robust method overcorrects the lowest level variance. At the second level, almost all effects are significant. The ML estimations give much larger deviations from the nominal coverage than the robust estimations. Again, we observe that having larger groups does not improve the situation. Robust standard errors are better than the asymptotic standard errors. For the symmetric

Table 6
Effects of the three conditions on the coverage of the 95% confidence interval for the random effects (first p-value for ML-estimation; second for robust estimation)

	E_0	U_0	U_1
No. of groups			
Chi^2	0.5736/0.0000*	0.1587/0.0000*	0.0429/0.0000*
Uniform	0.6441/0.0000*	0.0000*/0.0000*	0.0000*/0.0000*
Laplace	0.0297/0.7779	0.4137/0.0000*	0.0051/0.0000*
Group size			
Chi^2	0.0000*/0.0000*	0.0000*/0.0000*	0.0000*/0.0003*
Uniform	0.0000*/0.0000*	0.0000*/0.0000*	0.0000*/0.0000*
Laplace	0.0186/0.0000*	0.0000*/0.0000*	0.0000*/0.0045
ICC			
Chi^2	1.000/1.000	0.0000*/0.0012*	0.0000*/0.0000*
Uniform	1.000/1.000	0.0000*/0.0063	0.0000*/0.0000*
Laplace	1.000/1.000	0.0000*/1.000	0.0000*/0.8172

*sign.

uniform and Laplace distributed residuals the robust method appears to produce satisfactory confidence intervals. However, for the extremely skewed chi-squared residuals the resulting confidence intervals are largely biased, and only begin to approach their nominal coverage at the largest sample of groups ($NG = 100$) used in this simulation.

6. Summary and discussion

Non-normal distributed residual errors on the second (group) level of a multilevel regression model appear to have little or no effect on the estimates of the fixed effects. The estimates of the regression coefficients are unbiased, and both the ML and the robust standard errors are accurate. There is no advantage here in using robust standard errors. This corresponds to the general belief that ML estimation methods are generally robust (cf. Eliason, 1993).

Non-normal distributed residual errors on the second (group) level of a multilevel regression model do have an effect on the estimates of the parameters in the random part of the model. The estimates of the variances are unbiased, but the standard errors are not always accurate. At the lowest level, the ML standard errors are accurate, while the robust standard errors are overcorrected. The standard errors for the second-level variances are inaccurate for the uniform and Laplace distribution, and highly accurate for the chi-squared distribution. The robust errors tend to do better than the ML standard errors. If the distribution of the residuals is non-normal but symmetric (the uniform and Laplace distribution) the robust standard errors appear to work reasonably well. With the skewed chi-square residuals, all estimated confidence intervals are unacceptable. For chi-squared residuals, ML estimation produces for the 95% confidence interval for the parameters in the random part at the second level a coverage of only 66% and

Table 7
Effect of the number of groups and the group size on the coverage of the 95% confidence interval of the random effects (first the p-value for the ML estimation; second for the robust estimation)

	E_0	U_0	U_1
No. of groups			
30			
Chi^2	0.9487/0.9866*	0.6537*/0.8128*	0.6501*/0.8007*
Uniform	0.9506/0.9840*	0.9544/0.9613	0.9562/0.9660
Laplace	0.9553/0.9884*	0.8347*/0.8997*	0.8141*/0.8937*
50			
Chi^2	0.9539/0.9903*	0.6701*/0.8734*	0.6471*/0.8506*
Uniform	0.9530/0.9889*	0.9686/0.9790*	0.9657/0.9781*
Laplace	0.9456/0.9879*	0.8353*/0.9369	0.8278*/0.9221*
100			
Chi^2	0.9534/0.9933*	0.6659*/0.9217*	0.6308*/0.9059*
Uniform	0.9473/0.9922*	0.9760/0.9884*	0.9764/0.9887*
Laplace	0.9493/0.9903*	0.8444*/0.9621	0.8339*/0.9587
Group size			
5			
Chi^2	0.9373/0.9819*	0.7784*/0.9019*	0.7540*/0.8648*
Uniform	0.9380/0.9813*	0.9489/0.9700	0.9431/0.9646
Laplace	0.9441/0.9820*	0.8847*/0.9442	0.8587*/0.9203*
30			
Chi^2	0.9630/0.9937*	0.6219*/0.8582*	0.6032*/0.8500*
Uniform	0.9559/0.9908*	0.9717/0.9771*	0.9750*/0.9829*
Laplace	0.9528/0.9917*	0.8181*/0.9266	0.8187*/0.9330
50			
Chi^2	0.9557/0.9947*	0.5893*/0.8478*	0.5708*/0.8423*
Uniform	0.9570/0.9930*	0.9784*/0.9817*	0.9802*/0.9853*
Laplace	0.9533/0.9930*	0.8117*/0.9279	0.7984*/0.9211*
ICC			
0.10			
Chi^2	0.9520/0.9899*	0.7123*/0.8786*	0.6913*/0.8669*
Uniform	0.9501/0.9882*	0.9582/0.9719	0.9567/0.9714
Laplace	0.9499/0.9887*	0.8564*/0.9321	0.8452*/0.9279
0.20			
Chi^2	0.9521/0.9901*	0.6572*/0.8706*	0.6334*/0.8494*
Uniform	0.9501/0.9883*	0.9684/0.9772*	0.9678/0.9789*
Laplace	0.9502/0.9890*	0.8353*/0.9351	0.8213*/0.9250*
0.30			
Chi^2	0.9519/0.9902*	0.6201*/0.8588*	0.6032*/0.8408*
Uniform	0.9507/0.9886*	0.9723/0.9797*	0.9739/0.9824*
Laplace	0.9501/0.9890*	0.8227*/0.9314	0.8092*/0.9216*

*sign.

64%, compared to 87% and 85% for robust estimation. These results mean that when the group level residuals are skewed, neither the ML nor the robust estimation of the group level standard errors can be trusted. In the case of robust estimation, only having a very large number of groups (> 100) can compensate this, at the expense of having overcorrected standard errors at the lowest level.

In general we conclude that using ML methods for the analysis of multilevel data with non-normally distributed group level residual errors only causes problems when one is interested in the significance or in the confidence intervals of the variance terms at the second level. In that case, robust standard errors are more accurate. If the residuals have a non-normal but symmetric distribution, robust standard errors work generally well. If the distribution is markedly skewed, robust standard errors lead to confidence intervals that approach their nominal level only when the number of groups is large: at least 100. Raudenbush and Bryk (2002) suggest that comparing the asymptotic standard errors calculated by the ML method to the robust standard errors is a way to appraise the possible effect of model mis-specification, in addition to other methods such as inspecting residuals and formal tests. Hox (2002) extends this suggestion to model mis-specifications including violation of important assumptions. Used in this way, robust standard errors become an indicator for possible misspecification of the model or its assumptions. If the robust standard errors are much different from the asymptotic standard errors, this should be interpreted as a warning sign that some important assumption is violated. Clearly, the recommended action is not to simply rely on the robust standard errors to deal with the misspecification. Our simulation indicates that unless the number of groups is very large, the robust standard errors are not up to that task. Rather, the reasons for the discrepancy must be diagnosed and resolved.

If the residuals follow a markedly skewed distribution which cannot be resolved by altering the model or transforming variables, robust standard errors do not solve the problem. A different approach that merits the analysts' attention is the non-parametric bootstrap (cf. Carpenter et al., 1999; Hox, 2002) or a more general approach that allows non-normal distributions for the random effects.

7. Uncited reference

Bryk et al., 1996.

References

Afshartous, D., 1995. Determination of sample size for multilevel model design. Unpublished paper. 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, Bath, UK.

Browne, W.J., Draper, D., 2000. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Comput. Statist. 15, 391–420.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models. Sage, Newbury Park, CA.

Bryk, A.S., Raudenbush, S.W., Congdon, R.T., 1996. HLM. Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L programs. Scientific Software International, Chicago.

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

Carpenter, J., Goldstein, H., Rasbash, J., 1999. A non-parametric bootstrap for multilevel models. Multilevel Modelling Newslett. 11, 1, 2–5.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Mahwah, NJ.

Eliason, S.R., 1993. Maximum Likelihood Estimation. Sage, Newbury Park, CA.
Evans, M., Hastings, N., Peacock, B., 1993. Statistical Distributions. Wiley, New York.
Goldstein, H., 1995. Multilevel Statistical Models. Edward Arnold, London; Halsted, New York.
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. Amer. J. Epidemiol. 149, 876–883.
Hox, J.J., 1998. Multilevel modeling: when and why. In: Balderjahn, I., Mathar, R., Schader, M. (Eds.), Classification, Data Analysis, and Data Highways. Springer, New York, pp. 147–154.
Hox, J.J., 2002. Multilevel Analysis, Techniques and Applications. Lawrence Erlbaum Associates, Mahwah, NJ.
Hox, J.J., Maas, C.J.M., 2001. The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. Struct. Equation Modeling 8, 157–174.
Huber, P.J., 1967. The behavior of maximum likelihood estimates under non-standard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, pp. 221–233.
Kish, L., 1965. Survey Sampling. Wiley, New York.
Kreft, I., de Leeuw, J., 1998. Introducing Multilevel Modeling. Sage, Newbury Park, CA.
Leeuw, J.de., Kreft, I.G.G., 1986. Random coefficient models for multilevel analysis. J. Ed. Statist. 11, 57–85.
Longford, N.T., 1993. Random Coefficient Models. Clarendon Press, Oxford.
Maas, C.J.M., Hox, J.J., 2003. Sufficient sample sizes for multilevel modeling. Department of Methodology and Statistics, Utrecht University, NL, submitted for publication.
Muthén, B., Satorra, A., 1995. Complex sample data in structural equation modeling. In: Marsden, P.V. (Ed.), Sociological Methodology. Blackwell, Oxford, pp. 267–316.
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. Multilevel Models Project. University of London, London.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models, 2nd Edition. Sage, Thousand Oaks, CA.
Raudenbush, S., Bryk, A., Cheong, Y.F., Congdon, R., 2000. HLM 5. Hierarchical Linear and Nonlinear Modeling. Scientific Software International, Chicago.
Scott Long, J., Ervin, L.H., 2000. Using heteroscedasticity consistent standard errors in the linear regression model. Amer. Statist. 54, 217–224.
Snijders, T.A.B., Bosker, R., 1999. Multilevel analysis. An Introduction to Basic and Advanced Multilevel Modeling. Sage, Thousand Oaks, CA.
Tabachnick, B.G., Fidell, L.S., 1996. Using Multivariate Statistics. HarperCollins Publishers Inc., New York.
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, Leiden.
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.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.