Offprint

The Influence of Misspecification of the Heteroscedasticity on Multilevel Regression Parameter and Standard Error Estimates

Elly J.H. Korendijk, Cora J.M. Maas, Mirjam Moerbeek, and Peter G.M. Van der Heijden

Utrecht University, The Netherlands

Abstract. Like in ordinary regression models, in multilevel analysis, homoscedasticity of the residual variances is an assumption that is mostly unchecked. However, in experimental research, the residual variance component at level two may differ in the experimental and the control condition, leading to heteroscedastic second level variances. Using a simulation study, the consequences of ignoring second level heteroscedasticity on the estimation of the fixed and random parameters and their standard errors was investigated. It was found that the standard error of the second level variance is underestimated, but that the estimated fixed parameters of the independent variables, the first level variance and their standard errors are mostly unbiased.

Keywords: heteroscedasticity, homoscedasticity, heterogeneous variances, multilevel analysis, cluster randomized trials, simulation

Experimental research in social, educational and medical science often includes data with a so-called nested structure. Examples are children nested in villages and pupils nested in classes (Ausems, Mesters, Van Breukelen, & de Vries, 2002; Sommer et al., 1986). For ethical and practical reasons in these situations, groups and not single subjects are often assigned to different experimental conditions, that is, a number of groups receive treatment and the rest of the groups do not. Such designs are often referred to as *cluster randomized* trials. A nonignorable issue for designs like these is that subjects within groups may not be independent, leading to erroneous conclusions when ordinary regression analysis or analysis of (co)variances (AN[C]OVA) is used. Multilevel regression analysis corrects for this dependency and allows the treatment effect to be estimated corrected for covariates. When strong predictive covariates are taken into account, multilevel analysis has a high power to detect the treatment effect and gives precise estimates of the relevant parameters (Raudenbush, 1997).

The assumptions underlying the multilevel linear regression are similar to the assumptions in ordinary multiple regression analysis: linear relationships, homoscedasticity, and normally distributed residuals. For ordinary regression analysis, it is known that mild violations of these assumptions hardly affect the accuracy of the parameter estimates or standard errors. Thus, as long as the sample sizes are not too small, ordinary multiple regression analysis can be regarded as a robust method of analysis (Tabachnick & Fidell, 2001). For the multilevel regression model, it is shown by Maas and Hox (2003) that violations of the normality assumption at the second level have little or no effect on the parameter estimates and the standard errors of the fixed effects and of the lower level variance. Only the standard errors of the variances at the second level are inaccurate.

In ordinary regression analysis, in the case of severe violations, a variety of statistical methods for correcting heteroscedasticity are available (Scott Long & Ervin, 2000). In multilevel regression, the assumption of homoscedastic residuals relates to all levels. According to most authors (Bryk & Raudenbush, 1992; Goldstein, 1995; Snijders & Bosker, 1999), ignoring violation of the homoscedasticity assumption, will not only give incorrect values for the residual variances, but can also lead to biased standard errors and a worse model fit. It is strongly advocated to check for heteroscedasticity and model it if found, in order to get correct estimates for the residual variances and a better model fit. However, there are also a few examples showing that ignoring heteroscedastic variances hardly affects parameter estimates or the standard errors associated with these parameters (Roberts & Roberts, 2005; Snijders & Bosker, 1999, p. 112). Because violation of the assumption of homoscedastic variances is likely to occur in experimental research due to treatment itself and therapist effects (Roberts & Roberts, 2005), a systematic study investigating the effect of this violation is important and, therefore, addressed in this paper.

The main question to be answered in this paper is whether or not ignoring heteroscedasticity affects the parameter and the associated standard error estimates. Related questions are whether or not the severity of this effect is influenced by group size, intraclass correlation (ICC), or the magnitude of the difference between residual variances in the different treatment conditions. The first two conditions are chosen because research has shown that the number of groups (Maas & Hox, 2005) and the ICC (Goldstein, 1995) sometimes affect the parameter estimates. Further, it is reasonable to expect that, if heteroscedasticity affects the estimates, the magnitude of the bias of the estimates depends on the magnitude of the variance ratio. In the sections to follow, first the specification of a multilevel regression model and possible heteroscedasticity are described in more detail. Then, in the method section, the specifications of the conditions of the simulation are given, followed by a section in which the results are presented. The report ends with a discussion and conclusion.

The Multilevel Regression Model and Misspecification of Heteroscedasticity

Data from an experiment usually consist of information from J groups ($j = 1, 2, \ldots, J$) with a number of subjects n_j in each group. Individual i in group j has a continuous outcome variable Y_{ij}, one or more explanatory variables X_{ij} at the subject level, one or more explanatory variables Z_j at the group level, and a treatment indicator T_j. Modeling data like this, gives a separate regression equation for each group:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \beta_2 T_j + \beta_3 Z_j + e_{ij}. \quad (1)$$

in which β_1, β_2, and β_3 are the regression coefficients for X_{ij}, T_j and Z_j respectively and e_{ij} is the residual for person i in group j. The residuals are assumed to be normally distributed with zero mean and variance σ_e^2.

If the regression coefficients, β_1, β_2, and β_3 are homogeneous over groups, group differences are reflected by the regression coefficient β_{0j} (the intercept). This variation is modeled by:

$$\beta_{0j} = \gamma_0 + u_{0j}. \quad (2)$$

The subscript j in this equation shows that for each group an intercept is estimated, which consists of a term that is not group specific (γ_0) and a group specific term for the residual (u_{0j}). These second level residuals u_{0j} are assumed to be normally distributed with zero mean and variance σ_u^2. In general, it is assumed that σ_u^2 is the same for all groups:

$$\mathrm{VAR}(u_{0j}) = \sigma_{uj}^2 = \sigma_u^2. \quad (3)$$

This is the assumption of homoscedasticity. Substitution of (2) in (1) and assuming homoscedasticity gives:

$$Y_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 T_j + \gamma_3 Z_j + u_{0j} + e_{ij}. \quad (4)$$

Note that in this equation the β's are replaced by γ's. In multilevel models γ is the usual notation for fixed regression coefficients to differentiate them from β's that are allowed to differ between groups or subjects (compare equation [2]).

As argued in the previous section, in experimental research the homoscedasticity assumption may be violated because of the treatment itself and therapist effects. Therefore we assume that the heteroscedasticity depends on the treatment variable. In the simple experimental setting in which T is dichotomous with $T_j = 1$ for a group in the experimental condition and $T_j = 0$ for a group in the control condition, the intercept variance at the second level can be modeled by:

$$\sigma_{uj}^2 = \tau_0^2 + 2\tau_{01} T_j. \quad (5)$$

In equation (5), τ_0^2 is the variance in the control condition and $\tau_0^2 + 2\tau_{01} T_j$ the variance in the experimental condition (Snijders & Bosker, 1999). If the variance in the treatment condition is smaller than in the control condition, τ_{01} is negative, and it is positive if the variance in the treatment condition is larger than in the control condition. Analogous to equation (5) the residual term at the group level can be expressed as:

$$u_{0j} = u_{0j}^* + u_{1j} T_j, \quad (6)$$

in which u_{0j}^* is the residual for the control condition and $u_{1j} T_j$ the extra treatment dependent residual. Substitution of (2) and (6) in (1) and rearrangement gives:

$$Y_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 T_j + \gamma_3 Z_j + u_{0j}^* + u_{1j} T_j + e_{ij}. \quad (7)$$

In this specification of the multilevel model, the heteroscedastic variance at the second level is taken into account.

However, most researchers applying multilevel analysis use equation (4) whether the assumption of homoscedasticity is violated or not, and, hence, they misspecify the model if heteroscedasticity is present. In the current paper, the consequences of this misspecification are investigated in a simulation study.

Method

The MLwiN software (Rasbash, Steele, Browne, & Prosser, 2004) was used for the simulation and estimation. The latter was executed by using the restricted maximum likelihood method because in comparison with the full maximum likelihood method it produces less biased estimates of the variance components (Longford, 1993).

In the simulations, a cluster randomized design was

used, where half of the groups were assigned to the treatment condition and half of the groups to the control condition. Using equation (7; repeated here)

$$Y_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 T_j + \gamma_3 Z_j + u_{0j}^* + u_{1j} T_j + e_{ij} \quad \text{(7 repeated)}$$

heteroscedastic data were generated, with one explanatory variable at the subject level, X_{ij}, one explanatory variable at the group level, Z_j, and a treatment indicator variable T_j. The data were analyzed by using equation (4; repeated here):

$$Y_{ij} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 T_j + \gamma_3 Z_j + u_{0j} + e_{ij} \quad \text{(4 repeated)}$$

that is, assuming homoscedasticity in order to investigate the consequence of ignoring the heteroscedasticity.

Throughout all simulations, the group size was held constant at five. In applied social and behavioral sciences, such small group sizes are not uncommon, although in general group sizes will be larger. Maas and Hox (2005) have shown that even group sizes of five give correct parameter estimates and standard errors when all other conditions are satisfactory. When the results of the simulation study show that the parameter estimates are unbiased when group sizes are as small as five and the assumption of homoscedasticity is violated, they will be unbiased in larger groups as well.

Three conditions were varied in the simulation: (1) the number of groups (30, 60, 100, 200); (2) the intraclass correlation (ICC: 0.1, 0.2, 0.3); and (3) the variance ratio between the experimental and control group (1:2, 1:3). The specification for the number of groups was based on the knowledge of the influence of this variable on the standard errors of the estimates (Maas & Hox, 2005); 50 groups or less lead to underestimated standard errors of the second level variance. In order to exclude too small a number of groups as the cause of biased parameter estimates or biased standard errors, one specification for the number of groups was chosen to be as large as 200 (100 in each treatment condition). In practice, having 100 groups in each treatment condition is not very realistic. More realistic are 15, 30, or 50 groups in each treatment condition.

The ICC values span the customary range of ICC values as used in other simulation studies (Maas & Hox, 2003, 2005). The variance ratios were chosen as reasonable but well detectable differences.

There are $4 \times 3 \times 2 = 24$ combinations of the different conditions. For each combination 3,000 simulated data sets were generated, which results in a huge power. $\alpha = 0.001$ was therefore used as a criterion for significance and a Bonferroni correction (α divided by the number of tested parameters) was applied when more parameters were tested simultaneously.

For the regression coefficients, 1.00 was chosen for the intercept, and 0.3 for the three regression slopes (medium effect size; see Cohen, 1988; Maas & Hox, 2005). The first level variance σ_e^2 was fixed at 1.00. In the homoscedastic model, the value of the second level variance follows from the specification of the ICC and the first level variance by:

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}. \quad (8)$$

Given equation (8), and the specifications of the ICC's and the first level variance, the following values for the second level variance were used: 0.1111, 0.2500 and 0.4286 (in the homoscedastic model). These variance terms can be considered to represent the mean second level variance of the two different treatment conditions:

$$\sigma_u^2 = \frac{\tau_0^2 + (\tau_0^2 + 2\tau_{01} T_j)}{2}. \quad (9)$$

For determining the accurateness of the *parameter estimates* the relative bias was used, defined as the estimated parameter value divided by the true parameter value (*relative bias* = $\hat{\theta}/\theta$). In the ideal situation the estimated parameter equals the true parameter value and hence the relative bias is equal to one. By means of a *t*-test, the hypothesis "The relative bias equals 1" was tested for each parameter. If this overall *t*-test is significant, the estimate is biased. In order to determine in which condition the estimate was biased, *t*-tests were performed per condition. In case of more than one significant biased estimate within a condition (number of groups, ICC, or variance ratio), an ANOVA was executed to test the hypothesis that the bias within each condition is the same. If this hypothesis was rejected, the ANOVA was followed by post hoc pair wise comparisons, to see which estimates differ significantly.

In order to determine the accuracy of the *standard errors*, first the 95% confidence intervals around the estimated parameters in each simulated data set were established (Goldstein, 1995). Next, the frequency that the true parameter is in this 95% interval was counted. The coverage is the percentage of this count. The hypothesis "The coverage equals 95%" was tested by means of a confidence interval around the 95% point estimate. When the coverage of an estimated parameter is outside the confidence interval the standard error is biased. A bias upward, that is overestimated standard errors, leads to a relatively large confidence interval and, hence, more *chance* of containing the true parameter value that results in a larger coverage. A bias downward, that is, underestimated standard errors, results in coverage smaller than the ideal 95%.

In case of more than one significantly biased estimate within a condition (number of groups, ICC, or variance ratio), a χ^2 test was executed to answer the question whether or not the bias within a condition is the same.

Results

Convergence and Inadmissible Solutions

The estimation procedure converged in all simulations. In MLwiN it is possible that – especially when the true value is close to zero – the estimation procedure results in nega-

tive variance estimates. These estimates are called *inadmissible solutions*, because negative variances do not exist and are artifacts of the estimation process. In data analyzing practice, it is, therefore, conventional usage to set these negative variance estimates to zero. In the current simulation study, these negative variances were not recoded, because that would have led to values closer to the real parameter value and, therefore, to a too conservative test. In 12 of the 24 conditions, negative values with respect to estimation of the second level variance were found. In 2 conditions the percentage of negative second level variance estimates was almost 10%, in the other conditions less than 3%. 92% of the negative estimations occurred in the condition when the true parameter value equals 0.1111, which is close to zero.

Relative Bias

For assessment of the parameter estimates, first the relative bias (*relative bias* = $\hat{\theta}/\theta$) was calculated for each parameter. *t*-tests, performed on the mean relative biases to test the hypothesis of bias equal to 1, yielded only a significant bias for the first level variance. In order to determine in which condition the estimated first level variance is biased, *t*-tests per condition were performed. Three significant results were found. The first significant result was found when the number of groups is 30 ($t = -3.828$, $df = 17999$, $p = .00013$). The relative bias in this condition equals .9964, which is also the absolute value of the mean estimated parameter, since the true value of the parameter is 1. Though significant, the difference between the mean estimated parameter and the true parameter value is very small and, therefore, not relevant. The other two significant results were found within the condition of the variance ratio ($t = -3.948$, $df = 35999$, $p = .00008$ for variance ration 1:2 and $t = -3903$, $df = 35999$, $p = .00010$ for variance ratio 1:3). The estimated values, which equal the mean relative bias as explained above, are the same for both conditions. Both values are .9982 which is a small, irrelevant underestimation.

Coverage

In order to investigate the influence of the number of groups, the ICC and the variance ratio on the estimation of the standard errors of the parameters, the coverages per parameter per condition were calculated. Confidence intervals were established around the coverages to test the hypothesis that the coverage equals 95%. As explained in the Method section, an α of 0.001 was used as a criterion for significance and a Bonferroni correction was applied when more parameters were tested simultaneously. For the two random parameters, the confidence interval equals .9369 < CI < .9631 and for the four fixed parameters .9361 < CI < .9639.

Though some coverages significantly deviate from 95%, the deviances in almost all situations are very small and, therefore, not relevant. With respect to the fixed effects, Table 1, Table 2, and Table 3 show that the coverages associated with the intercept are overestimated in almost all conditions and that for one condition the coverage of the treatment effect is underestimated. The largest coverage

Table 1. Effects of the number of groups on the coverages for the fixed and random parameters

Number of groups	200	100	60	30
Fixed parameters				
Intercept	.9705*	.9697*	.9648*	.9647*
First-level predictor	.9535	.9476	.9443	.9418
Second-level predictor	.9491	.9471	.9459	.9434
Treatment effect	.9422	.9445	.9434	.9325*
Random Parameters				
First-level variance	.9461	.9453	.9441	.9443
Second-level variance	.9338**	.9336**	.9226**	.9090**

*significant outside the interval .9361 < CI < .9639 for the fixed parameters; **significant: outside the interval .9369 < CI < .9631 for the random parameters.

Table 2. Effects of the magnitude of the ICC on the coverages for the fixed and random parameters

ICC	.30	.20	.10
Fixed parameters			
Intercept	.9735*	.9690*	.9597
First-level predictor	.9459	.9477	.9468
Second-level predictor	.9460	.9481	.9451
Treatment effect	.9406	.9414	.9400
Random Parameters			
First-level variance	.9447	.9451	.9450
Second-level variance	.9211**	.9245**	.9286**

*significant outside the interval .9361 < CI < .9639 for the fixed parameters; **significant: outside the interval .9369 < CI < .9631 for the random parameters.

Table 3. Effects of the variance ratio on the coverages for the fixed and random parameters

Variance ratio	1:2	1:3
Fixed parameters		
Intercept	.9631	.9718*
First-level predictor	.9468	.9468
Second-level predictor	.9454	.9474
Treatment effect	.9411	.9403
Random Parameters		
First-level variance	.9450	.9449
Second-level variance	.9269**	.9226**

*significant outside the interval .9361 < CI < .9639 for the fixed parameters; **significant: outside the interval .9369 < CI < .9631 for the random parameters.

with respect to the intercept, 97.35%, was found in the condition where the ICC is .30. Even in this worst case, the estimation is less than 1% outside the confidence interval. To test whether or not the significant coverages for the intercept differ within the condition, χ^2 tests were performed. No significant results were found. The underestimation of the treatment effect in the condition where the number of groups is 30, is 0.29% and, hence, irrelevant.

With respect to the random effects, the results show that the coverages for the first level variance are never deviating from 95%. The coverages for the second level variance are significantly underestimated in all conditions. χ^2 tests revealed no differences of the coverages within the condition "ICC," nor within the condition "variance ratio." The maximum underestimation is 1.5% and does not seem relevant.

However, in the condition "number of groups," the χ^2 test is significant ($\chi^2 = 107.064$, $df = 3$, $p < .0002$), meaning that the coverages within this condition are not equal. As can be seen in Table 1, in the condition number of groups is 30, the coverage is only 90.90%. This is 2.7% below the lower bound of the confidence interval and was judged to be a nonnegligible underestimation. Increasing the number of groups results in less underestimation, though even with 200 groups the coverage is still outside the confidence interval.

Discussion and Conclusion

First the results are summarized. *t*-tests on the relative biases of the estimated parameters showed that almost all fixed and random parameters are unbiased. Three times the first level variance is significantly underestimated. However, the deviations from the true parameter are less than 0.01, so not relevant at all.

Most standard errors are also unbiased, as was shown by the established confidence intervals around the coverages. Some exceptions on this were seen: The standard errors of the intercept are overestimated in almost all conditions, the standard error of the treatment effect is underestimated in the condition with a small number of groups, and the standard errors of the second level variance are underestimated in all conditions. What are the implications of these findings?

First, it should be noted that the single term estimated by the homoscedastic model for the second level variance hides the existing difference between the variance in the control condition and the variance in the experimental condition. Although the total second level variance is estimated without bias, the heteroscedasticity of the model is misspecified. The underestimation of the standard errors of the second level variance are therefore of minor importance, because they belong to the already incorrectly modeled variance term, which for most researchers is a nuisance parameter.

The overestimation of the standard errors associated with the intercept also causes no problems. Not only are most researchers not interested in the intercept, the overestimation is also very small and therefore not relevant.

The parameters of main interest to most researchers – the parameters associated with the treatment effect and the covariates – are all estimated without bias. Almost all standard errors belonging to these parameters are also unbiased. Only the standard error associated with the treatment effect is once underestimated, but the underestimation is too small to be of any importance.

Based on these results, it can be concluded that having unequal variances in the experimental and the control condition, and analyzing the data ignoring this existing heteroscedasticity, does not lead to erroneous conclusions about treatment effects and the effect of the covariates. When it is accepted that the estimated second level variance gives no insight into possible existing heteroscedasticity and the magnitude of the ratio, and when one is mainly interested in the effect of the treatment and the covariates, using a homoscedastic model in case of heteroscedastic second level variance is proven to be robust.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research, grant NWO-400-05-097.

References

Ausems, M.M.P.H., Mesters, I., Van Breukelen, G., & de Vries, H. (2002). Short-term effects of a randomized computer-based out-of-school smoking prevention trial aimed at elementary schoolchildren. *Preventive Medicine, 34,* 581–589.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* London: Erlbaum.

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

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

Maas, C.J.M., & Hox, J.J. (2003). The influence of violation of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis, 46,* 427–440.

Maas, C.J.M., & Hox, J.J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1,* 86–92.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MLwiN version 2.0.* London: Multilevel Models Project, University of London.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster-randomized trials. *Psychological Methods, 2,* 173–185.

Roberts, C., & Roberts, S.A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials, 2,* 152–162.

Scott Long J., & Ervin, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54,* 217–224.

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

Sommer, A., Djunaedi, E., Loeden, A.A., Tarwotjo, I., West, K.P. Jr., & Tilden, R. (1986). Impact of vitamin A supplementation on childhood mortality. A randomized control Community trial. *The Lancet, 8491,* 1169–1173.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics.* Boston: Allyn and Bacon.

Elly Korendijk

Department of Methodology and Statistics
Faculty of Social and Behavioral Sciences
Utrecht University
PO Box 80.140
NL-3508 TC Utrecht
The Netherlands
E-mail e.j.h.korendijk@uu.nl