Analyzing Clustered Data with OLS Regression: The Effect of a Hierarchical Data Structure

Daniel M. McNeish
University of Maryland, College Park

A previous study by Mundfrom and Schultz (2002) has shown that coefficient estimates between Ordinary Least Squares (OLS) regression and Hierarchical Linear Models (HLM) for clustered data are, for all intents and purposes, equivalent. However, when comparing these two methods, the standard error estimates are the true cause for concern. A simulation study is conducted to demonstrate that the standard error estimates between OLS and HLM are quite different for clustered data, and to show that the differential standard errors leads to inflated Type I error rates using OLS under such conditions.

Frequently, in social science research, observations have a hierarchical structure (Raudenbush & Bryk, 2002). Students are nested within classrooms, workers are nested within organizations, or patients are nested within hospitals. While this hierarchal structure may seem obvious, it has a large impact when analyzing the results of data from such a structure.

Methods comprising the general linear model, such as multiple linear regression (MLR), are extremely versatile for analyzing and modeling data. However, a key assumption of these methods is that observations are independent (Cohen, Cohen, West, & Aiken, 1983). When data are clustered and come from a hierarchical structure, this assumption is often violated. Observations coming from the same higher level unit are more like one another than observations from a separate higher level unit. For example, in education, students in classroom A are more like one another on reading ability than students in classroom B since students in the same class receive instruction from the same teacher who has particular strengths and weaknesses. This creates some dependence in the data for which MLR cannot account.

To determine whether MLR may be unsuitable for the data, a quantity called the design effect (DEFF) can be calculated. DEFF is a measure that addresses how the sampling variability is affected as a result of departures from simple random sampling (Hahs-Vaughn, 2005; Kish, 1965). DEFF values greater than 1.0 indicate that the sampling variability is greater as a result of a hierarchical structure, with DEFF values greater than 2.0 generally being considered to indicate that the hierarchical nature of the data should be taken into account to avoid underestimated standard errors (Satorra & Muthén, 1995).

Should DEFF values be above 2.0, the statistical literature has developed methods for addressing data that come from a hierarchical structure and can account for the dependence among observations. One commonly implemented method in social science research has many names and acronyms, but is often referred to as hierarchical linear modeling (HLM, used in this paper), multilevel modeling, or mixed effect models (Bickel, 2007; Raudenbush & Bryk, 2002; Singer & Willet, 2003).

Research Questions

Although HLM is specifically designed to address hierarchical structures, practitioners are not always aware of HLM, do not always employ HLM as part of their data analysis, or may be unaware that the data have a hierarchical structure and utilize MLR instead. Mundfrom and Schultz (2002) showed that it is not always problematic for some model estimates. In their simulation study, coefficient estimates were compared when hierarchical data were analyzed using MLR and HLM for a wide range of intraclass correlations (ICCs), a measure that shows how homogenous data are within clusters (Raudenbush & Bryk, 2002). Mundfrom and Schultz found that the coefficient estimates (or fixed effects as they are deemed in HLM) between MLR and HLM did not differ greatly, even with ICC values as high as 0.95 (i.e., ICC values are capped at 1.00 so values close to 1.00 suggest that the clustering of the data has an enormous impact).

While Mundfrom and Schultz (2002) did not find a difference between coefficient estimates under HLM and MLR for different ICC levels, the coefficient estimates are not the primary concern when the independence assumption within MLR is violated. Rather, the standard error estimates present a more pressing issue. When data have a hierarchical structure, observations within clusters are more related to one another than to other observations. This results in the reduction of the effective sample size, which directly impacts the standard error estimates (Bickel, 2007). For example, imagine a situation in which

there are 10 clusters of 10 units each. In the extreme case where responses within a cluster are identical (i.e., the ICC is 1.00 meaning that the cluster affiliation has a great deal of impact on the outcome), variability within the clusters would be zero and the only variability would be between clusters. In HLM, the effective sample size would be 10, whereas MLR would utilize a sample size of 100 because observations are assumed to be independent. Since standard error estimates are a function of sample size, MLR and HLM would result in quite different standard error estimates in this case.

These standard errors are vitally important to a practitioner's interpretation of his or her model. Statistical significance is typically determined by a *t*-test which takes a ratio of the coefficient estimate with its standard error. If standard errors estimates are too small, *t* values will be inflated, which leads to higher rejection rates than otherwise stipulated by the nominal type I error rate. Even though this may seem trivial to some, the consequence could result in the erroneous declaration of a parameter as statistically significant that could affect conceptualizations of the phenomenon of interest.

In addition to the standard errors, the degrees of freedom used in inferential tests would be different between MLR and HLM as well. In MLR, the degrees of freedom for a single-parameter *t*-test are *N-p-1* where *N* is the total sample size and *p* is the number of predictor variables plus the intercept. In HLM, degree of freedom calculation is more tenuous with a few competing methods being commonly employed such as contain, between, residual, Satterthwaite (Satterthwaite, 1941), and Kenward-Roger (Kenward & Roger, 1997). For a more thorough discussion of these methods, interested readers are referred to Schaalje, McBride, and Fellingham (2002) or Table 56.7 of the SAS 9.2 User's Manual.

While Mundfrom and Schultz (2002) do mention the differential estimation of standard errors between HLM and MLR, they did not present results regarding standard error estimates between the two models. This study will employ a simulation design to replicate findings in Mundfrom and Schultz that coefficient estimates are quite similar between MLR and HLM, and will also extend their study by demonstrating that standard error estimates between MLR and HLM will diverge when data have a hierarchical structure. The intention is to illustrate that although the choice of model will not greatly affect the coefficient estimates, the standard error estimates will be quite different since MLR will not address the dependency between observations.

Simulation Design

The data in this study were generated by a two-level random intercepts model using only a single level-1 predictor variable. The exact model is as follows,

$$y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

Where γ_{00} and γ_{10} are the fixed effect coefficients for the intercept and slope, respectively, u_{0j} is the random effect for the intercept, r_{ij} is the residual, X_{1ij} is a level-1 predictor variable, and y_{ij} is the outcome variable that is assumed to be continuous. There is no random effect for the slope nor are there level-2 or cross-level predictors in this model.

Factors Not Manipulated

Certain factors in this design will not be manipulated and will remain constant for all replications. Sample size both at level-2 (the cluster level) and level-1 (within-clusters) will be held constant throughout all replications. Three-hundred clusters each with 50 units will be generated for a total sample size of 15,000 per replication. One-thousand replications will be run for each ICC condition for a total of 4,000 replications. Each simulated dataset will be analyzed twice, once using MLR with OLS estimation and again using HLM with restricted maximum likelihood (REML) estimation.

HLM literature suggests that the 30 clusters with 30 units is the minimum recommended sample size for estimating fixed effects using multilevel models with continuous outcome variables since they are estimated with maximum likelihood, an asymptotic estimation method that requires large sample sizes (Kreft, 1996; Maas & Hox, 2005). Much larger values were chosen to ensure that any findings from this study are attributable to manipulated factors rather than inadequate sample sizes.

The population parameter values for both fixed effects ($\gamma_{00}$and γ_{10}) will also remain unchanged throughout the replications. γ_{00} will be set at 15 while γ_{10} will be set at 0.35. Values for the level-1 predictor variable will be generated randomly from a normal distribution with a mean of 20 and a standard deviation of 5.

Factors Manipulated

Two factors in the design will be manipulated. First, the ICC will be altered. The ICC explains how homogenous observations are within clusters with values near 1.00 indicating a complete homogeneity and values near 0 indicating complete independence. Four conditions of ICC will be investigated: 0.00, 0.10, 0.17, and 0.50. The null value of 0.00 is chosen since this represents the condition where the clustering of the data has no effect and the independence assumption of MLR is met. Standard error estimates should be very close between MLR and HLM under this condition and differ only due to sampling variability and small differences between OLS and maximum likelihood estimation.

The values of 0.10 and 0.17 were selected based on typical ICC values mentioned in the literature. Muthén (1994) suggested that typical ICC values for HLM are between 0.00 and 0.50 while Bliese (2000) suggested that ICC values between 0.05 and 0.20 are more common. The first value of 0.10 was chosen as a low-moderate ICC while 0.17 was chosen at a high-moderate ICC. The final value of 0.50 was chosen to represent the upper extreme of commonly observed ICC values.

The second factor to be manipulated is the statistical model, either MLR or HLM. Each generated dataset will be analyzed with each model. Fixed effect coefficient estimates and their standard errors will be compared.

Outcome Measures

Three outcome measures will be collected to address the proposed research questions. The first outcome measure will address the relative bias of the fixed effect coefficient estimates. True population values for both the intercept and slope were selected as 15 and 0.35, respectively. Relative bias was calculated as follows:

$$\text{relative bias} = \left(\frac{\frac{1}{1000}\sum_{k=1}^{1000}(\hat{\gamma}_k - \gamma_{(\text{true})})}{\gamma_{(\text{true})}} \right)$$

where *k* is an indicator for the replication. This will be done for both slope and intercept for each ICC condition in each model. This formula will give the percentage by which the coefficient estimates differ from the true value across all replications. Positive values indicate that the coefficients were overestimated (i.e., too high) while negative values indicate that the coefficients were underestimated (i.e., too low).

For the question addressing standard errors, relative bias is not a viable outcome measure since standard errors cannot be directly set in a simulation design. The second outcome measure will be a ratio of the average estimated standard error in each condition compared to the standard deviation of the coefficient estimate across all 1,000 replications for each condition. The average estimated standard error is an approximation of the sampling variability of a coefficient while the standard deviation of the coefficient estimates across replications is one way to assess the actual variability of the coefficients within a simulation design. The ratio is deemed standard error efficacy and is calculated by $\text{SE Efficacy} = \overline{\text{SE}}/_{\text{SD}}$. The closer this value is the 1.00, the closer the estimates are to one another. When the estimates are close, it shows that the particular estimation method does reasonably well estimating the sampling variability of the coefficient (i.e., the standard error). If the efficacy is far from 1.00, this provides evidence that estimation of the standard error for the particular estimation method is biased. Based on criteria set forth by Hoogland and Boomsma (1998), standard error efficacies less than 0.90 indicate underestimated standard errors while values above 1.10 indicate overestimated standard errors.

The third outcome measure will be the 95% confidence interval coverage. This measure records the proportion of replications in which the population value for the coefficient appears in the estimated confidence interval for a particular parameter that addresses both the bias of the coefficient estimate and the bias of the standard error simultaneously. If the coefficient estimate is biased, then the interval will be centered around a biased value. If the standard error is biased, then the width of the interval will be too short if standard errors are underestimated and too wide if standard errors are overestimated. Based on

criteria set forth by Bradley (1978), 95% confidence interval coverage rates below 0.925 or above 0.975 are considered problematic for an α of 0.05.

Results

Relative Bias

Overall, the coefficient estimates were almost identical to the true values for each model in each ICC condition for both the intercept and the slope. These results are similar to those found by Mundfrom and Schultz (2002). The absolute value of the relative bias never exceeds 0.2% for any set of conditions simulated in this study.

Standard Error Efficacy

Since the slope predictor was not specified in the generating model to be dependent on the hierarchical structure (i.e., there was not random effect for slope), only standard error efficacy results for the intercept will be presented. As expected, the standard error efficacy for the intercept was very similar between estimation methods when the ICC was 0.0 (HLM = 1.00, MLR = 1.00 for the intercept). This represents a situation where the clustering effect in the data is null so either statistical model is appropriate. However, even though the HLM standard error estimates stayed fairly consistent throughout the ICC conditions with the standard error efficacy never falling below 0.97, the MLR standard error efficacy for the intercept steadily fell as the ICC increased. For an ICC of 0.50, the standard error efficacy was 0.52 meaning that the standard error was estimated to be about half of the true sampling variability. Figure 1 graphically depicts the consistency of the standard error efficacy when using HLM and the downward trend in the standard error efficacy when using MLR with non-zero ICCs.

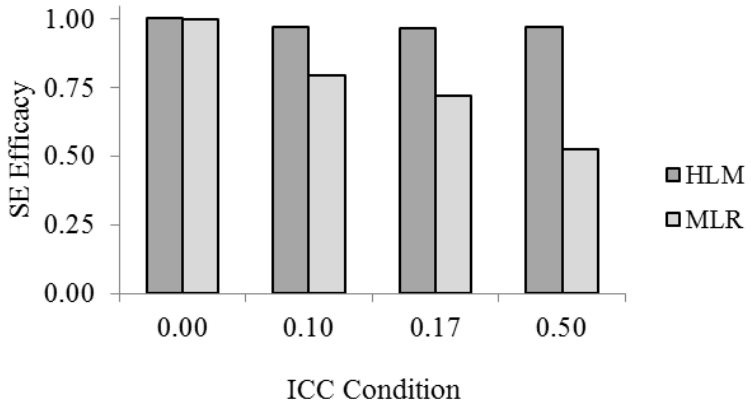

Figure 1. Standard error efficacy for intercept by ICC condition.

95% Confidence Interval Coverage

Table 1 shows the coverage of the 95% confidence interval for the intercept across ICC conditions as well as the estimated DEFF, which is calculated by $DEFF = 1 + (\text{average cluster size} - 1) * ICC$. Again, values for the slope are not shown since they were not affected by the hierarchical structure in the data generation model. For HLM, the values across all ICCs are very close to the nominal 0.95 level. When the ICC is 0, MLR provides comparable coverage to HLM. However, when the ICC is non-zero, the coverage rates for the intercept can be seen to decrease steadily and well below acceptable criteria of 0.925 outlined in Bradley (1978). Additionally, the higher the ICC, the further the coverage rate decreases. Although this measure includes information about both coefficient bias and standard error bias, as shown above and in Mundfrom and Schultz (2002), coefficient estimates exhibit virtually no bias. Thus, the 95% confidence coverage results would indicate that standard error estimates are inappropriate for MLR with non-zero ICCs and that Type I error rates would be inflated in such scenarios with Type I error rates approaching 0.30 in the 0.50 ICC condition. With

Table 1. Design Effect and 95% confidence interval coverage for intercept by ICC and model

	ICC			
Model	0.00	0.10	0.17	0.50
HLM	0.95	0.96	0.95	0.95
MLR	0.95	0.88	0.84	0.71
DEFF	0.00	5.90	9.33	25.50

regard to the design effect, all non-zero ICC conditions had DEFF values over 2.0. Congruent with Satorra and Muthén (1995), simulation conditions with DEFF values greater than 2.0 led to narrow 95% confidence coverage intervals with MLR, which indicates that the standard errors are underestimated and methods that account for clustering are necessary.

Discussion

As found previously in Mundfrom and Schultz (2002), regardless of model choice, coefficient estimates will be unbiased when data come from a hierarchical structure and the model is properly specified. However, standard error estimates are a larger concern when choosing a method to analyze data that may come from a hierarchical structure. When clustered data are analyzed via OLS regression, standard errors will be underestimated since they do not account for the dependency in observations.

Practitioners must be sure to verify the source of their data before proceeding with OLS regression. Although it may seem obvious that clustered data would benefit from being analyzed with a method designed for clustered data, it may not always be clear that one has meaningfully clustered data. Even when the clustering is minimal, the standard error estimates may be inappropriately underestimated with OLS regression. The tenability of the independence assumption (along with the other assumptions) must be rigorously tested to ensure that inferences are appropriate. Otherwise, MLR may be used inappropriately and Type I error rates will consequently be inflated. If one has data that may have a hierarchical structure, the degree of dependence can be tested by running an unconditional HLM model where no predictors are included in the model, but a random effect for the intercept is included. From this model, an estimate of the ICC can be calculated and used to address whether HLM may be warranted. A common rule of thumb from Hox (1998) is to use HLM if the ICC is greater than 0.05. Additionally, once the ICC is calculated, design effects can be estimated for each level-1 predictor to help adjudicate whether clustered data methods should be implemented. Of course, if the hierarchical nature of the data set itself is of substantive interest (e.g., the research question calls for modeling slope variability using one or more level-2 predictors), then HLM may be appropriate regardless of the magnitude of the ICC or design effect, since level-1 slopes could vary significantly regardless of the ICC value.

As in Mundfrom and Schultz (2002), MLR did not appear to show any issues when estimating coefficients for the slope and the intercept. So, if researchers are interested in estimating the coefficients only for descriptive purposes or for use in prediction, MLR seems appropriate for such a situation even if data are clustered. However, it should be noted that statistical significance tests may be untrustworthy in such scenarios.

This study also had certain limitations as well. The data generation was kept very simple. A model with only one level-1 predictor, zero level-2 predictors, and only one random effect is unlikely to be representative of models being conducted in practice. Further, significance tests of the intercept are not always of interest. However, the intercept was selected to simplify presentation of the results and a similar relation would hold for other predictors in the model.

Further, MLR and HLM are not the only two possible approaches that can be used with regression models, which may oversimplify the problem. Design-based methods, such as generalized estimating equations (Liang & Zeger, 1986) or survey methods such as Taylor series linearization (Heeringa, West, & Berglund, 2010), can be used to account for clustering when model assumptions have been violated (i.e. the independence assumption in MLR) without having to specify the clustering in the model with random effects as found with fixed effects models that include each individual cluster in the model as a predictor (Allison, 2005). Nearly all general statistical software programs support these types of modeling techniques.

References

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Raleigh, NC: SAS Institute.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: Guilford Press.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein, & S. W. Koszlowski (Eds.), *Multilevel theory, research and*

methods in organizations: Foundations, extensions, and new directions (pp. 349-381). San Francisco: Jossey-Bass.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education*, *73*, 221-248.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329-367.

Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). Berlin, Germany: Springer.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983-997.

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

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished manuscript, California State University, Los Angeles .

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13-22.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 86-92.

Mundfrom, D., & Schultz, M. (2002). A Monte Carlo simulation comparing parameter estimates from multiple linear regression and hierarchical linear modeling. *Multiple Linear Regression Viewpoints*, *28*, 18-21.

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 276-398.

Raudenbush, S. A., & Byrk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Satorra, A., & Muthen, B. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267-316.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*, 309-316.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 512-524.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Send correspondence to: Daniel M. McNeish
University of Maryland, College Park
Email: dmcneish@umd.edu