March, 2005

Using OLS Models to Facilitate Student Understanding of HLM Models

John Fraas

Using OLS Models to Facilitate Student Understanding of HLM Models

David O. Newman

Cleveland State University

Russell Brown

Cleveland Municipal School District

John W. Fraas

Ashland University

Joshua Bagakas

Cleveland State University

James A. Salzman

Cleveland State University

Isadore Newman

The University of Akron

Paper presented at the annual meeting of the Eastern Educational Research Association

Sarasota, Florida

March 3-6, 2005

Abstract

It is important for graduate students to understand the various research questions their hierarchical linear models are actually testing. If the students construct variables and multiple linear regression models that produce nearly identical results to their hierarchical linear models, we believe their understanding of the direct relationship between a hierarchical linear model and its specific research question will be facilitated. We present three examples that illustrate this pedagogical approach.

Using OLS Models to Facilitate Student Understanding of HLM Models

Hierarchical Linear Modeling (HLM) has become an important analytical tool in a number of fields of study, include the field of education. This sophisticated technique allows the researcher to analyze data that has multiple levels. For example, in educational research one might want to investigate student level, class level and building level data. Because of this capacity, many of the educational journal articles published during the last decade have used this technique. We believe it is important for graduate students who use HLM to understand what is being modeled by their hierarchical linear models. If graduate students are not sure what is being modeled, the chance of committing Type VI errors, which occurs when the model does not match the research question, is increased (Newman, Deitchman, Burkholder, Sanders, & Ervin, 1976; Newman, Fraas, Newman, & Brown, 2002).

Many graduate students have had previous exposure to multiple linear regression models that are estimated with the ordinary least squares technique. We trust they have learned to construct models to reflect specific research questions [one should refer to Newman, McNeil, and Kelly (1996) for an excellent discussion of how to construct models to test specific research questions]. We believe the student understanding of an introduction to HLM can be facilitated if the professor demonstrates how a multiple linear regression model must be constructed to reflect the corresponding hierarchical linear model. Such an exercise may facilitate the graduate students' understanding, and possibly the professor's understanding as well, of what the hierarchical linear model is analyzing. This understanding will better enable the students to assess whether the model reflects the research question of interest. The remaining sections of this paper provide

three illustrations of how to construct the variables and design a multiple linear regression model or models that are capable of testing the same research question as a given hierarchical linear model.

Data File and Variables Used with the OLS Models

The hypothetical data used for our examples were assumed to have been generated from a repeated-measures design with individuals nested within time. For simplicity we limited the data set to 20 individuals with each individual measured three times. One of the important aspects of this exercise is to have the students clearly understand how these data must be entered into a data file and how the variables are created for use in the OLS models. For the three illustrations included in this paper, it should be noted that we used the SPSS® computer software to create our variables and analyze our multiple linear regression models.

The first characteristic that is important to note regarding our data file is that it contained 60 cases, that is, 60 rows. Each person required three rows in the data file because three scores were recorded for each person. Thus the three scores for Person 1 were entered in the first three rows under the variable name scores. The scores for the other 19 individuals were entered in the same manner in the column containing the scores variable.

The second variable, which was named time, contained the same three values for each person. The numbers 1, 2, and 3 indicated the time period from which the scores were recorded (i.e., Time1, Time2, and Time3). The next 20 variables, which were labeled person1 through person20, were person variables. One should refer to Newman, et al. (1996), Pedhazur (1977), and Williams (1977) for excellent presentations of the

concept of person variables (vectors). Each person variable contains only zero and one values. For the person1 variable, a value of one was entered in row 1 through row 3. The value of one indicates that the first three scores in the data set were recorded for the first person. Since the remaining 57 scores in the data file were not recorded for the first person, the value of zero was entered for each of the remaining 57 rows. The values for the other 19 person variables were entered in the same fashion.

The last 20 variables entered in the data file were designed to represent the interaction effect between the time and person variables. Each of these 20 interaction variables were formed by multiplying the time variable by a given person variable. These interaction variables were labeled time*person1 through time*person20.

To summarize, the data file we used in conjunction with the OLS models contained a total of 42 variables and 60 cases. The scores variable contained the three scores recorded for each of the 20 individuals. The time variable contained values 1, 2, and 3, which indicated the time period in which each score was recorded. The 20 person variables, which contained zero and one values, indicated the person for which each score was recorded. And the 20 interaction variables were computed to represent the interaction effect between time and individuals.

The OLS and HLM Models Used in Conjunction with Research Question 1

To illustrate how the construction and analysis of an OLS model can facilitate a graduate student's understanding of a hierarchical linear model we began by posing a rather straightforward research question. The first question posed, which was labeled Research Question 1, was as follows:

Research Question 1: Do the scores reflect a linear trend?

This question requires the researcher to construct a model that estimates the linear change in the scores over time.

The OLS model that contains this requirement is as follows (Note: The scores were converted to Z scores):

Z_Score = a0U + a1Time + Error (OLS Model 1)

It is important for the students to note two characteristics of this model. First, this model *ignores* the information regarding person differences. That is, the model does not allow for different intercept points for different individuals. Second, the model does not include the interaction variables. Thus, this model requires the slopes of the trends of scores to be the same for all 20 individuals. We doubt that students would understand these explicit assumptions the researchers are making when using this model unless they understand the concept of person vectors and the interaction of these person vectors with the time variable.

The HLM Model 1 used to test Research Question 1 is as follows:

LEVEL 1 MODEL

$$Z\_SCORE_{ti} = \pi_{0i} + \pi_{1i}(TIME_{ti}) + e_{ti}$$

LEVEL 2 MODEL

$$\pi_{0i} = \beta_{00}$$

$$\pi_{1i} = \beta_{10}$$

Where:

1. Level 1 = within person effects
2. Level 2 = between person effects.

In this model the error term attributable to persons on both the intercept and slope are fixed.

Table 1 contains the results produced by the OLS Model 1 and HLM Model 1. A review of Table 1 by the graduate students revealed to them that the coefficients, standard error, and significance levels produced by OLS Model 1 and HLM Model 1 for the time variable were nearly identical. These nearly identical results indicated to the graduate students that HLM Model 1 does not allow (a) different intercept points for different individuals or (b) different trends (slopes) for individuals. We believe these basic assumptions, which are incorporated into HLM Model 1, are more readily understood through the construction of the data set used in conjunction with OLS Model 1 and the variables not included, as well as those included, in that model.

Insert Table 1 about here

The OLS and HLM Models Used in Conjunction with Research Question 2

To illustrate that an HLM model can explicitly assume the intercept points and slopes of the individual trend lines to be different, we posed a second research question. The second question posed, which was labeled Research Question 2, was as follows:

Research Question 2: Do the trends of the scores differ across individuals?

This question requires that the OLS models and the HLM model used to test it must allow the intercept points and the slopes of the trend lines of the scores for individuals to differ.

Two OLS models must be designed to test this research question. The first OLS model, which allows different intercept points and slopes, would contain 19 of the 20 person variables. Students should be reminded that the series of 20 person variables are linearly dependent. Thus, one of the person variables cannot be entered into the OLS

model. It should also be noted that the horizontal summation of the interaction variables produces a vector identical to the time vector. Thus, if the time variable is included in the model, only 19 of the 20 interaction variables can be included in the model. The person variable that we did not enter into the OLS model was person20. Thus, OLS Model 2a was constructed as follows:

Z_Score = a0U + a1Time + a2Person1 … a19Person19 +

a20Person1*Time…a39Person19*Time + error (OLS Model 2a)

The second OLS model restricts the slopes to be equal. This model, which was labeled OLS Model 2b, contained the time variable and the same 19 person variables that were contained in OLS Model 2a. Thus, OLS Model 2b was constructed as follows:

Z_Score = a0U + a1Time + a2Person1 … a20Person19 + error (OLS Model 2b)

HLM Model 2, which would be used to test Research Question 2 was designed and analyzed. The HML Model 2 was designed as follows:

LEVEL 1 MODEL

$$Z\_SCORE_{ti} = \pi_{0i} + \pi_{1i}(TIME_{ti}) + e_{ti}$$

LEVEL 2 MODEL

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

Where:

1. Level 1 = within person effects (Time)
2. Level 2 = between person effects (Persons)

Table 2 contains the results produced by OLS Model 2a, OLS Model 2b, and HLM Model 2. A review of the statistical test of the F test of the change in the R^2 value for OLS Model 2b and the R^2 value for Model 2a and the chi-square test of the time slope

value in HLM Model 2 revealed the slopes were invariant across individuals. That is, the interaction effects between the time and person variables were not significant.

Insert Table 2 about here

The OLS and HLM Models Used in Conjunction with Research Question 3

Since the interaction term was not significant, the researchers may continue their analysis by investigating another research question. This question, which was labeled Research Question 3, is as follows:

Research Question 3: Does a common trend in the scores exist allowing for differences in the initial scores?

The OLS model designed to test this research question would include the time variable and 19 of the 20 person variables. The OLS model was constructed as follows:

Z_Score = a0U + a1Time + a2Person1 … a20Person19 + error (OLS Model 3)

The students should note the interaction variables are not included in this OLS model because the research question assumes the trends of the scores (slopes) for the 20 individuals are equal.

The HLM model designed to test this question is as follows:

LEVEL 1 MODEL

$$Z\_SCORE_{ti} = \pi_{0i} + \pi_{1i}(TIME_{ti}) + e_{ti}$$

LEVEL 2 MODEL

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10}$$

Note the error term for the slope has been fixed, but the error for the intercepts has been allowed to vary by subject.

The results produced by OLS Model 3 and HLM Model 3 are listed in Table 3. When the students review the results listed in Table 3 they will realize that the coefficient, standard error, and statistical test values for the time variable are nearly identical.

Insert Table 3 about here

Summary

Our goal for this paper was to present a pedagogical method by which graduate students would develop a better introductory understanding of what research questions HLM models test. We believe that by constructing the variables used in OLS models and the OLS models themselves, students will have a clearer understanding of the type of research question HLM models are testing. Through this understanding of the need to ensure that the model is testing the research question of interest, students will be aware of the importance of avoiding Type VI errors, that is, the lack of congruency between the analytic technique employed by the researcher and the research question.

Reference

McNeil, K., Newman, I., Kelly, F. J. (1996). *Testing research hypotheses with the general linear model*. Carbondale, IL: Southern Illinois University Press.

Newman, I., Deitchman, R., Burkholder, J., Sanders., & Ervin, L. (1976). Type VI error: Inconsistency between the statistical procedure and the research question. *Multiple Linear Regression Viewpoints, 6*(4), 1-19.

Newman, I., Fraas, J., Newman, C. & Brown, R. (2002). Research practices that produce Type VI Errors. *Journal of Research in Education, 12* (1), 138-145.

Pedhazur, E. J. (1977). Coding subjects in repeated-measures designs. *Psychological Bulletin, 84*, 298-305.

Williams, J. D. (1977). A note on coding the subjects effect in treatments x subject designs. *Multiple Linear Regression Viewpoints, 8*(1), 32-35.

Table 1

Results of OLS Model 1 and HLM Model 1

OLS Model 1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.716	.176		-4.069	.000
	Time	-.207	.136	-.195	-1.517	.135

HLM Model 1

Fixed Effect	Coefficient	Standard Error	T-ratio	d.f.	Approx. P-value
For INTRCPT1, P0					
INTRCPT2, B00	-0.716145	0.176009	-4.069	58	0.000
For TIME slope, P1					
INTRCPT2, B10	-0.206760	0.136336	-1.517	58	0.135

Table 2

Results of OLS Model 2a, OLS Model 2b, and HLM Model 2

OLS Model 2a and OLS Model 2b

Model Summary [c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.962[a]	.926	.888	.29149	.926	24.433	20	39	.000
2	.974[b]	.949	.848	.33935	.023	.462	19	20	.951

HLM Model 2

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	R0	0.78768	0.62044	19	183.31805	0.000
TIME slope,	R1	0.03123	0.00098	19	12.01292	>.500
level-1,	E	0.29003	0.08412			

Table 3

Results of OLS Model 3 and HLM Model 3

OLS Model 3

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-.321	.192		-1.673	.102
	p1	1.220	.238	.308	5.128	.000
	p2	-.133	.238	-.034	-.559	.579
	p3	-.356	.238	-.090	-1.496	.143
	p4	1.273	.238	.321	5.349	.000
	p7	.352	.238	.089	1.479	.147
	p8	-1.265	.238	-.319	-5.314	.000
	p9	.148	.238	.037	.622	.537
	p10	.867	.238	.218	3.641	.001
	p11	-.303	.238	-.076	-1.272	.211
	p12	-.734	.238	-.185	-3.082	.004
	p14	-.162	.238	-.041	-.682	.499
	p24	.719	.238	.181	3.021	.004
	p25	-.347	.238	-.087	-1.458	.153
	p26	-1.238	.238	-.312	-5.203	.000
	p27	.350	.238	.088	1.472	.149
	p28	-1.275	.238	-.321	-5.357	.000
	p29	-1.504	.238	-.379	-6.318	.000
	p30	-1.159	.238	-.292	-4.870	.000
	p31	-.221	.238	-.056	-.931	.358
	Time	-.207	.046	-.195	-4.486	.000

a. Dependent Variable: Score

HLM Model 3

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	d.f.	Approx. P-value
For INTRCPT1, P0 INTRCPT2, B00	-0.716145	0.185814	-3.854	19	0.001
For TIME slope, P1 INTRCPT2, B10	-0.206760	0.046387	-4.457	19	0.000