22 ELEMENTS OF HIERARCHICAL LINEAR REGRESSION MODELS

Once you know hierarchies exist, you see them everywhere.[1]

In this chapter we consider models that are increasingly popular in social, behavioral, educational, and medical research. They go by various names, such as **multilevel models** (MLM), **hierarchical linear models** (HLM), **mixed-effect models** (MEM), **random-effects models** (REM), **random coefficient regression models** (RCRM), **growth curve models** (GCM), and **covariance components models** (CCM). Because they share several common features, and for brevity of discussion, we will call all such models HLM. Where necessary we will point out the special features of the various models.

At the outset, it should be noted that the subject of HLM is vast and mathematically demanding. There are several specialized books written on this subject.[2] If you surf the Internet for the general term "hierarchical linear models", you will find a plethora of articles, both theoretical and empirical, on various areas of social and other sciences.[3] Several software packages specially written for HLM are cited in the following references.[4] For our purpose we will use *Stata*, which also has a routine to estimate such models.

This chapter introduces the bare bones of the nature and significance of HLMs. With an extended example, we show how such models are formulated, estimated, and interpreted. Knowledge of HLMs will also help the reader understand some of the limitations of the linear regression models that we have discussed in this text.

1 Kreft, I. and De Leeuw, J., *Introducing Multilevel Modeling*, p. 1. Sage Publications, California, 2007.

2 See, for example, Luke, D. S., *Multilevel Modeling*, Sage Publications, California, 2004; Twisk, J. W. R., *Applied Multilevel Analysis*, Cambridge University Press, Cambridge, 2006; Hox, J. J., *Multilevel Analysis: Techniques and Applications*, 2nd edn, Routledge, 2010; Bickel, R., *Multilevel Analysis for Applied Research: It's Just Regression!*, Guilford Press, 2007; Gelman, A. and Hill, J., *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, Cambridge, 2007; Byrk, A. S. and Raundenbush, S. W., *Hierarchical Linear Models*, Sage Publications, California, 1992. For more advanced discussion with applications, see Rabe-Hesketh, S. and Skrondal, A., *Multilevel and Longitudinal Modeling Using Stata: Vol 1: Continuous Responses*, 3rd edn, Stata Press, 2012. Volume 2 deals with Categorical Responses, Counts and Survival. Finally, there is Hox, J., Multilevel modeling: when and why, in I. Balderjahn, R. Mathar, and M. Schader (eds.), *Classification, Data Analysis, and Data Highways*, Springer-Verlag, Berlin, 1998, pp. 147–54..

3 For an interesting application of HLM involving President Obama's 2008 election, see http://www.elecdem.eu/media/universityofexter/elecdem/pdfs/intanbulwkspjan2012/Hierarc.

4 For a list of multilevel modeling software packages, see Luke, *op cit*., p. 74.

22.1 The basic idea of HLM

Data for research often have a hierarchical, or multilevel, structure in the sense that **micro-level**, or lower level, data are often embedded in **macro-level**, or higher level, data. An often-cited example is from the field of education, where students are grouped in classes, classes are grouped in schools, schools in school districts, school districts in counties, counties in states, and so on. Thus, each lower unit of observation is part of a chain of successively higher level of data. In situations such as this, analyzing data at the lower, or micro-, level without taking into account the *nested*, or *clustered*, nature of the data can lead to erroneous conclusions. Thus, it is very important to keep in mind the *context* in which the data are collected. Therefore, HLM models are often called **contextual models**.

The primary goal of HLM is to predict the value of some micro-level dependent variable (i.e. regressand) as a function of other micro-level predictors (or *regressors*) *as well as some predictors at the macro level*. However, the macro-level predictors are not introduced directly into the regression model, as in the classical linear regression model, but in an indirect way, as explained below. For discussion purposes, we will call analysis at the micro-level as **Level 1** analysis, and that at the macro-level, **Level 2** analysis. To keep the discussion simple, we will consider only Level 1 and Level 2 analysis, but the analysis can be easily extended to more than two levels. As you would expect, if we consider several levels of data, the analysis becomes increasingly complicated.

Hierarchical data are typically found in survey data. Some surveys are very extensive, with layers of information collected on a variety of subjects. In the USA there are about 30 major surveys on a variety of topics.[5] In this chapter, we will discuss one such survey, the **National Educational Longitudinal Survey** (**NELS**).

22.2 NELS

NELS is a longitudinal survey that follows eighth-grade students as they transit into and out of high school. The objective of the survey is to observe changes in students' lives during adolescence and the role that school plays in promoting growth and positive life choices. The study began in 1988 when the students were in the eighth grade and followed them through grade 12; they also followed students who dropped out of high school as well. In the first three waves, data were collected on achievement tests in mathematics, reading, social studies, and science.

The baseline 1988 survey included 24,599 students, one parent of each student respondent, school principals (about 1,302), and teachers (about 5,000). The vast amount of data collected makes it possible to analyze it at various levels, such as parent, teacher, and school as well as to conduct analysis by region, race/ethnicity, public vs. private school, and the like. It is only when you examine the actual data that you will appreciate the wealth of the data. You can analyze it at several levels, depending on your interest and computing facilities.

In the data, the student-level, or Level 1, variables are socio-economic status of students (*SES*), the number of hours of homework done per week (*Homework*), student's race (white coded as 1, and non-white coded as 0), parents' education level

5 For an overview of these surveys, see Vartanian, T. P., *Secondary Data Analysis*, *Pocket Guides to Social Work Research Methods*, Oxford University Press, Oxford, 2011.

(*Parented*), and class size, measured by student/teacher ratio (*ratio*), . The macro-level, or Level 2, variables are school id (*schid*), education sector (public schools coded as 1 and private schools coded as 0) (*Public*), percentage of ethnic minority students in school (% minorities), geographic region of school (Northeast, North Central, South and West), represented by dummy variables, and composition of school (urban, suburban, and rural) (*Urban*), also represented by dummy variables.

The dependent variable in our analysis is the score on a math test, which is a Level 1 variable.

22.3 Analysis of NELS data

Since this chapter is but an introduction to HLM, we will consider only a simple two-level HLM. For higher-level analysis, the reader can consult the references.

We use the NELS-88 data, but to keep the discussion manageable, we use a smaller subset of the data, namely 10 randomly selected schools from 1003 schools. Information is collected on 260 students from these 10 schools, which is a tiny fraction of the original 21,580 students in the full data set. Again, this is strictly for illustrative purposes.[6]

Our objective in this mini-study is to find out the impact of the number of hours of homework (a Level 1 variable) on the score on a math test, which is Level 1 analysis. Later we will add more Level 1 regressors to the model. The macro variable chosen for the initial analysis is the school ID (*schid*) which is a Level 2 variable. Of course, we could choose any of the macro variables mentioned above. Not only that, we could have more than one Level 2 variable added to the model. The data set used in the analysis is available as **Table 22.1** on the companion website.[7]

OLS analysis of NELS data: the naïve model

To appreciate HLM, we first consider the relationship between the score on a math test (*Math*) and the number of hours of math homework (*homework*) using OLS. In OLS, we will consider all 260 students regardless of their school ID (*schid*) in studying the relationship between the two variables. This may be called the **pooled regression**. We first consider a **naïve** or **null** regression model in which there is no regressor. That is we estimate:

$$Math_i = B_1 + u_i \tag{22.1}$$

where $Math$ = score on a math test, u is the error term, and i is the ith student, and where we assume the error term follows the usual (classical) OLS assumptions, in particular the assumption that $u_i \sim N(0,\sigma^2)$, that is they are identically and independently distributed as a normal variate with zero mean and constant a variance. In short, they are NIID. The intercept, B_1, in this model is assumed to be fixed, for its value is assumed the same across all schools and individuals. Hence, we can call (22.1) a **fixed coefficient model**. If we estimate this regression, what we obtain is the average math score of all 260 students regardless of their school affiliation.

6 Exercise 22.1 provides data for 519 students in 23 schools.

7 We analyze the same data that was used by Kreft *et al.*, so that the reader can compare our analysis with theirs. Later they extend their analysis to 519 students in 23 schools.

Since our data are clustered into 10 schools, it is quite likely that one or more assumption of the (normal) classical linear regression model may not hold true. To allow for this possibility, we estimate regression (22.1) with *Stata*'s *robust standard errors* option.[8] The results are shown in Table 22.2.

Table 22.2 Naïve regression with robust standard error option.

. regress math, robust

Linear regression		
Number of obs	=	260
F(0, 259)	=	0.00
Prob > F	=	.
R-squared	=	0.0000
Root MSE	=	11.136

		Robust				
math	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[95% Conf. Interval]**	
_cons	51.3	.6906026	74.28	0.000	49.94009	52.65991

Since there is no regressor in this model, the *_cons* (for constant) simply represents the average math score of 260 students regardless of their school affiliation in the sample. This average score is statistically "highly significant".

Although the results are based on the *robust* standard errors, they may not be reliable, for the robust option in the present case neglects the fact that our data are clustered into school districts (*schid*). It is very likely that observations within a cluster (i.e. school here) are correlated: It is difficult to maintain that the math scores of students in the same school are uncorrelated. In fact, students in the same school tend to have test scores that are correlated, since they all study in the same environment and probably have similar backgrounds. This correlation is called the **clustering problem** or the **Moulton Problem**, named after Brent Moulton, who published an influential paper on this subject.[9] Briefly, the Moulton problem arises because the standard regression techniques applied to hierarchical data often exaggerate the statistical significance of the estimated coefficients. As we will show shortly, the standard errors of the estimated coefficients are underestimated, thereby exaggerating the estimated t values.[10]

The standard error reported in Table 22.2 does not correct for the Moulton problem, even though it may correct for heteroscedasticity in the error term. One way to take into account the clustering problem is to use the **clustered standard errors**. These standard errors allow regression errors to be correlated within a cluster, *but*

8 The *robust* option in *Stata* uses the Huber–White sandwich estimator. Such standard errors try to take into account some of the violations of the classical linear regression assumptions, such as non-normality of regression errors, heteroscedasticity, and outliers in the data.

9 See Moulton, B. (1986) Random group effects and the precision of regression estimates, *Journal of Econometrics*, **32**, 385–97.

10 The reason for this is that with correlation among observations in a given school (or cluster), we actually have fewer independent observations than the actual number of observations in that school.

assume that the regression errors are uncorrelated across clusters. If we use the cluster option, we need not use the robust option, for the latter is implied in the former.[11]

Using *Stata*'s **cluster** option, we obtain the results in Table 22.3.

Table 22.3 Estimation of the naïve model with the clustered standard errors.

regress math, cluster (schid)

Linear regression

Number of obs	=	260
F(0, 9)	=	0.00
Prob > F	=	.
R-squared	=	0.0000
Root MSE	=	11.136

(Std. Err. adjusted for 10 clusters in schid)

math	**Coef.**	Robust **Std. Err.**	**t**	**P>|t|**	**[95% Conf. Interval]**	
_cons	51.3	3.402609	15.08	0.000	43.60276	58.99724

The command **cluster (schid)** is *Stata*'s command to use the clustered standard error option with the cluster variables, in this case a single Level 2 variable, *schid*. If you compare the results in Table 22.3 with those in Table 22.2, you will notice that the coefficient value remains the same in both cases, but the standard errors are vastly different. This suggests that the error term in the naïve model suffers from heteroscedasticity, autocorrelation, or other related problems. As a result, the OLS standard errors, even with the robust option, are severely underestimated.[12]

In the present case, even with the cluster option, the estimated coefficient is highly significant. However, this cannot be taken for granted in all situations.

22.4 HLM analysis of the naïve model

Instead of using OLS with the robust or the clustered standard error approach, we now consider HLM modeling of the test scores, which incorporates both Level 1 and Level 2 variables. The Level 2 variable we use is the school ID (*schid*).

The OLS intercept values obtained in Table 22.2 or Table 22.3 assume that the intercept value remains the same across the 10 schools. This is an unrealistic assumption. One way to find out if this is so is to introduce nine dummy variables to represent the 10 schools: Remember the rule that the number of dummy variables must be one less than the number of categories of the dummy variable, here the 10 different schools. Of course, if we had included all the 1003 schools in our sample, we would need 1002 dummies. Aside from the question of the power of the statistical

11 It may be noted that clustered standard errors are sometimes known as *Rogers standard errors*, since Rogers implemented them in *Stata* in 1993. See Rogers, W. (1993) sg17: Regression standard errors in clustered samples, *Stata Technical Bulletin*, **13**, 19–23.

12 If the number of clusters or groups (schools in the present case) is small relative to the overall sample size, the clustered standard errors could be somewhat larger than the OLS results.

tests[13] resulting from the analysis, this does not make much practical sense. In HLM, we do not estimate separate intercepts for the various schools. Instead, we assume that these intercepts are randomly distributed around a (grand) mean value with certain variance. More specifically, we assume

$$Math_{ij} = B_{1j} + u_{ij} \tag{22.2}$$

where $Math_{ij}$ = math score for student i in school j and B_{1j} = intercept value of school j, i going from 1 to 260 and j going from 1 to 10. We further assume that

$$B_{1j} = \gamma_1 + v_j \tag{22.3}$$

where v_j is the error term. The coefficient γ_1 represents the mean value of math score across all students and across all schools. We can call it the *grand mean*. The individual class mean math score varies around this grand mean. We assume the error term v_j has zero mean and constant variance.

Combining Eqs. (22.2) and (22.3), we obtain:

$$\begin{aligned} Math_{ij} &= \gamma_1 + v_j + u_{ij} \\ &= \gamma_1 + w_{ij} \end{aligned} \tag{22.4}$$

where

$$w_{ij} = v_j + u_{ij} \tag{22.5}$$

That is, the composite error term w_{ij} is the sum of school-specific error term v_j (the Level 2 error term) and student-specific error term u_{ij} (the Level 1 error term), or the regression error term. Assuming these errors are independently distributed, we obtain:

$$\sigma^2_{w_{ij}} = \sigma^2_j + \sigma^2_{ij} \tag{22.6}$$

That is, the total variance is the sum of the error variance due to the school effect and due to the individual student, or the usual regression error term.

If we take the ratio of the school-specific variance to the total variance, we obtain what is known as the **intra-class correlation coefficient** (**ICC**),[14] which is denoted by rho (= ρ)

$$\text{ICC} = \rho = \frac{\sigma^2_j}{\sigma^2_j + \sigma^2_{ij}} = \frac{\sigma^2_j}{w^2_{ij}} \tag{22.7}$$

It gives the proportion of the total variation in math scores that is attributable to differences among schools (i.e. clusters). In general, ICC is the proportion of the total variance that is between groups or clusters. A higher ICC means school differences account for a larger proportion of the total variance. To put it differently, a higher ICC means each additional member of a cluster provides less unique information. As a result, in cases of high ICC, a researcher would prefer to have more clusters with fewer members per cluster than to have more members in a small number of clusters.

13 The power of a test is the probability of rejecting the null hypothesis when it is false; it depends on the value of the parameters under the alternative hypotheses.

14 ICC is much different from the usual Pearson's coefficient of correlation. The former is correlation of observations within a cluster, whereas the latter is correlation between two variables. For example, a Pearson correlation coefficient of 0.3 may be considered small, but an ICC of 0.3 is considered quite large.

Once we estimate (22.4), we can easily obtain the value of ICC given in Eq. (22.7). For this purpose, we need to use statistical software specifically designed to estimate such models. Toward that end, we can use the **xtmixed** command in *Stata* 12.[15] The **xtmixed** procedure in *Stata* fits linear mixed models to hierarchical data. Mixed models contain both **fixed effects** and **random effects**. The fixed effect is given by the coefficient γ_1 and the random effect is given by w_{ij}.

The fixed effects are similar to the usual regression coefficients and are estimated directly. However, the random effects are not estimated directly but are obtained from their estimated variances and covariances. Random effects mean random intercepts, or random slopes, or both that take into account the clustered nature of the data. In estimating such models, the error term is usually assumed to be normally distributed.[16] The estimated coefficients involve one or more iterations, which are usually obtained by the Newton–Raphson interative procedure.

We first present the results of (22.4) (Table 22.4) and then comment on them.

Table 22.4 HLM regression results of model (22.4): random intercept but no regressor.

xtmixed math || schid:,variance
performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log likelihood = −937.38956
Iteration 1: log likelihood = −937.38956
Computing standard errors:

Mixed-effects ML regression	Number of obs	=	260
Group variable: schid	Number of groups	=	10
	Obs per group: min	=	20
	avg	=	26.0
	max	=	67
	Wald chi2(0)	=	.
Log likelihood = −937.38956	Prob > chi2	=	.

math	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**[95% Conf. Interval]**	
_cons	48.87206	1.835121	26.63	0.000	45.27529	52.46883

Random-effects Parameters	**Estimate**	**Std. Err.**	**[95% Conf. Interval]**	
schid: Identity				
var(_cons)	30.54173	14.49877	12.04512	77.44192
var(Residual)	72.23582	6.451525	60.63594	86.05481

LR test vs. linear regression: chibar2(01) = 115.35 Prob >= chibar2 = 0.0000

Note: The term “identity” means that these are random effects at Level 2 variable, *schid*, in the present case.
Note: The Wald statistic is not reported because there are no regressors in the model.

15 *Stata* has an alternative procedure, called **gllamm**, that can also estimate the mixed effects models.
16 If the assumption of normality and large samples are not met, the ML estimates are unbiased, but their standard errors are biased downward. On this see, Van Deer Leeden, R., Busing, F., and Meijer, E. (1997) *Applications of Bootstrap Methods to Two-level Models*. Paper presented at Multilevel Conference, Amsterdam, April 1–2. On bootstrap methods, See Chapter 23.

Before we discuss these results, let us look at *Stata*'s **xtmixed** command. The **xtmixed** command is followed by the regressand (*math* in our case), followed by two vertical lines, followed by the school ID (*schid*), which is the Level 2 variable, which is followed by options; here we use the option "variance", which tells *Stata* that we need the variances of the two error terms. Without this option, *Stata* produces standard deviations of the two error terms (i.e. the square root of the variance).

The output is divided into three parts. The first part gives general statistics, such as the number of observations, the number of groups (schools in the present case), the value of the likelihood function, and the Wald statistic, which like the R^2 in OLS, gives the overall fit of the model. In the present example, there are no regressors, so the value of Wald statistic is not reported. But in general, the Wald statistic has degrees of freedom and it follows the chi-square distribution.

The second part of the output gives information about the estimated coefficients, their standard errors and their z statistics (remember we are using the ML method with normal distribution), and their p values. Here the estimated value of the common intercept γ_1 is about 48.87, which is highly significant. The coefficient(s) reported in this part are fixed effects. But notice that this value is smaller than that obtained from the OLS regression with the robust or the clustered standard method. The standard error of the fixed coefficient is also different from that obtained in Tables 22.2 or 22.3.

The third part of the table gives the estimates of the error variances, σ_j^2 and σ_{ij}^2, their standard errors, and the 95% confidence intervals. Both these estimates are statistically significant.[17] Notice that the estimated σ_{ij}^2 of about 72.23 is much smaller than the estimate of the error variance given in Table 21.2 or Table 22.3, which is 124.01 ($=11.136^2$). What this suggests is that much of the latter error variance is accounted for by the introduction of the random intercept to the model, that is, by explicitly considering the impact of the Level 2 variable.

Let us examine the output in Table 22.4 further. The error variances, σ_j^2 and σ_{ij}^2 are, respectively, 30.54 and 72.23, giving the total variance of 102.77. Using Eq. (22.7), we obtain an ICC value of about 0.30. This value suggests that about 30% of the total variation in math scores is attributed to a Level 2 variable (school ID). This means that the math scores of students in a school are not independent, which violates a critical assumption of the classical linear regression model. To put it differently, in analyzing math scores we should not neglect the nested nature of our data.

This is further confirmed by the result of the likelhood ratio (LR) test,[18] whose value is given at the end of the output of Table 22.4. The LR test compares the random coefficient model with the fixed coefficient OLS regression. Since this test is significant, we can conclude that the random coefficient model is preferable to the fixed-coefficient OLS model.

Some software packages produce a statistic called **deviance**, which is a measure of judging the extent to which a model explains a set of data when parameter estimation is carried out by the method of maximum-likelihood (ML). It is computed as follows:

$$\text{Deviance} = -2lf \tag{22.8}$$

where lf is the log-likelihood function value. The smaller the value of the deviance, the better the model. For the naïve OLS model, the deviance value is $-2(-995.06) =$

17 If σ_j^2 is zero, there is no need to consider the random intercept model.

18 The LR test is discussed briefly in the appendix to Chapter 1.

1990.12. For the naïve HLM model, the deviance is –2(–937.38) = 1874.76. Therefore, on the basis of deviance, the HLM naïve model is preferable to the OLS naïve model.

More formally, if we use HLM instead of OLS, the deviance is reduced by about 115.36 (1990.12 – 1874.76), which is simply the value of the LR statistic given at the end of Table 22.4. And this LR value is highly statistically significant, for its *p* value is practically zero.[19]

Neglecting ICC can have serious consequences on Type I error, for the nominal and actual levels of significance can differ substantially. Assuming a nominal level of significance of 5%, a sample size of 50, and an ICC of 0.20, the actual level of significance is about 59%. Also, assuming a nominal level of significance of 5%, sample size of 50 and an ICC of 0.01, the actual level of significance is 11%.[20]

What all this suggests is that one should not neglect ICC in analyzing multilevel data.

22.5 OLS and HLM regressions with regressors

The naïve model is useful to bring out the importance of the ICC in analyzing clustered data. But in reality we use models that use one or more regressors. To keep things simple, let us introduce the number of hours of homework as an explanatory variable to explain the performance on a math test. Later on, we will add more regressors.

First, we consider an OLS regression:

$$Math_i = B_1 + B_2 Homework_i + u_i \qquad (22.9)$$

Again, note that we are pooling 260 observations to estimate this regression, without worrying about the Level 2 variable. The results are shown in Table 22.5.

We also present the results of Eq. (22.9) with clustered standard errors (Table 22.6).

As you would expect, there is a positive and statistically significant relationship between the grade on a math test and the number of hours of homework. But notice that the clustered standard errors are substantially higher than the OLS standard errors. Thus, it is important to take the structured nature of the data explicitly in the analysis.

Equation (22.9) is a fixed coefficient model, for it assumes that the regression coefficient is the same across all schools. This assumption may be as unrealistic as the assumption that the intercept remains the same across schools. Shortly, we will relax these assumptions with HLM modeling.

22.6 HLM model with random intercept but fixed slope coefficient

We now consider model (22.9), allowing for a random intercept but a fixed slope coefficient:

---

19 Kreft and De Leeuw suggest that one model has significant improvement over another model if the difference between their deviances is at least twice as large as the difference in the number of estimated parameters. See Kreft *et al*., p. 65.

20 For further details, see Barcikowski, R. S. (1981) Statistical power with group mean as a unit of analysis, *Journal of Educational Statistics*, **6**(3), 267–85.

Table 22.5 OLS regression of math grades on hours of homework.

.regress math,homework robust

Linear regression

	Number of obs	=	260
	F(1, 258)	=	88.65
	Prob > F	=	0.0000
	R-squared	=	0.2470
	Root MSE	=	9.6185

math	Coef.	Robust Std. Err	t	P>|t|	[95% Conf.	Interval]
homework	3.571856	.379369	9.42	0.000	2.824802	4.31891
_cons	44.07386	.9370938	47.03	0.000	42.22854	45.91918

Note: Root MSE is the standard error of the regression, that is, the square root of the error variance. The latter is therefore about 93.73. For this model, the log-likelihood statistic is −958.1770.

Table 22.6 OLS regression of math grades on hours of work with clustered standard errors.

. regress math homework, cluster(schid)

Linear regression

	Number of obs	=	260
	F(1, 9)	=	21.54
	Prob > F	=	0.0012
	R-squared	=	0.2470
	Root MSE	=	9.6815

(Std. Err. adjusted for 10 clusters in schid)

math	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
homework	3.571856	.7695854	4.64	0.001	1.830933	5.312779
_cons	44.07386	2.23222	19.74	0.000	39.02423	49.12349

$$Math_{ij} = B_{1j} + B_2 Homework_{ij} + u_{ij} \tag{22.10}$$

In this model, the intercept is random, but the slope coefficient is fixed (there is no subscript j on B_2.) Now instead of estimating separate intercept for each school, we postulate that the random intercept in Eq. (22.10) varies with the school ID, the Level 2 variable, as follows:

$$B_{1j} = \gamma_1 + \gamma_2 schid_j + v_j \tag{22.11}$$

where *schid* = school ID. Notice how the original intercept parameter, B_{1j}, now becomes the dependent variable in the regression (22.11). This is because we are now treating B_{1j} as a random variable.

What Eq. (22.11) states is that the random intercept is equal to the average intercept for all schools (= γ_1) and that it moves systematically with the school ID; each school may have special characteristics.

Combining Eqs. (22.10) and (22.11), we obtain:

$$\begin{aligned} Math_{ij} &= \gamma_1 + \gamma_2 schid_j + v_j + B_2 Homework_{ij} + u_{ij} \\ &= \gamma_1 + \gamma_2 schid_j + B_2 Homework_{ij} + (v_j + u_{ij}) \\ &= \gamma_1 + \gamma_2 schid_j + B_2 Homework_{ij} + w_{ij} \end{aligned} \tag{22.12}$$

where $w_{ij} = v_j + u_{ij}$, that is the composite error term w_{ij} is the sum of school-specific error term and the regression error term, which are assummed to be independent of each other. In this model the original intercept B_{1j} is not explicitly present, but it can be retrieved in the estimating procedure, as discussed below.

For this model the total, school-specific, and student specific variances are the same as in Eq. (22.6). The estimated values of these variances will enable us to estimate the ICC.

Using the **xtmixed** command of *Stata* 12, the regression results of Model (22.12) are as shown in Table 22.7 (compare this output with that given in Table 22.6).

Table 22.7 Results of regression (22.12): random intercept, constant slope.

xtmixed math homework || schid:,variance
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log likelihood = –921.32881
Iteration 1: log likelihood = –921.32881
Computing standard errors:

Mixed-effects ML regression	Number of obs	=	260
Group variable: schid	Number of groups	=	10
	Obs per group: min	=	20
	avg	=	26.0
	max	=	67
	Wald chi2(1)	=	34.37
Log likelihood = –921.32881	Prob > chi2	=	0.0000

math	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
homework	2.214345	.3777094	5.86	0.000	1.474048	2.954642
_cons	44.97838	1.724798	26.08	0.000	41.59784	48.35892

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
schid: Identity				
var(_cons)	22.50327	10.99337	8.638015	58.62426
var(Residual)	64.2578	5.741049	53.93568	76.55536

LR test vs. linear regression: chibar2(01) = 73.70 Prob >= chibar2 = 0.0000

Interpretation of results

As in Table 22.6, the first part of the table gives summary measures, such as the number of observations in the sample, the number of macro- or group variables (10 in the present case), the log-likelihood function, and the Wald (chi-square) statistic as a measure of the overall fit of the model. In the present case, the Wald statistic is highly significant, suggesting that the model we have used gives a good fit. The log-likelihood value is not particularly useful in the present case; it is useful if we are comparing two models.

The middle part of the table gives output that is quite similar to the usual OLS output, namely, the regression coefficients, their standard errors, their z (standard normal) values, and the 95% confidence intervals for the estimated coefficients. As you can see, the estimated coefficients are highly statistically significant.

The next part of the table is the special feature of HLM modeling. It gives the variance of the random intercept term (=22.50), the error variance (=64.25), from which we can compute the total variance (σ_w^2 = 86.75 = 22.50 + 64.25). From these numbers, we obtain:

$$ICC = \frac{22.50}{86.75} \approx 0.26 \tag{22.13}$$

That is, about 26% of the total variance in math scores is accounted for by differences in schools. This result, therefore, casts doubt on the OLS results given in Table 22.6. The LR test given at the end of Table 22.7 shows that the random intercept/constant slope model, which takes into account explicitly the Level 2 variable, school ID, is preferable to the OLS model.[21]

22.7 HLM with random intercept and random slope

Just as we allowed the intercept to be random, we can also allow the slope coefficient(s) to be random. Here again we can use multiplicative dummies to allow for random slope coefficients. But with several regressors, including multiplicative dummies will unnecessarily consume degrees of freedom. All this can be avoided if we use the random slope coefficient model, which again can be done much more economically with HLM. Toward that end, consider the following model:

$$Math_{ij} = B_{1j} + B_{2j} Homework_{ij} + u_{ij} \tag{22.14}$$

In this model, both intercept and slope coefficients are random, as they carry the j (Level 2) subscript.

We assume that the random intercept evolves as per (22.11) and the random slope evolves as follows:

$$B_{2j} = \lambda_1 + \lambda_2 schid_j + \omega_j \tag{22.15}$$

Substituting for B_{1j} and B_{2j}, we obtain:

$$\begin{aligned} Math_{ij} &= (\gamma_1 + \gamma_2 schid_j + \nu_j) + (\lambda_1 + \lambda_2 schid_j + \omega_j) Homework_{ij} + u_{ij} \\ &= \gamma_1 + \gamma_2 schid_j + \lambda_1 Homework_{ij} + \lambda_2 schid_j Homework_{ij} \\ &\quad + \omega_j Homework_{ij} + \nu_j + u_{ij} \\ &= \gamma_1 + \gamma_2 schid_j + \lambda_1 Homework_{ij} + \lambda_2 schid_j Homework_{ij} \\ &\quad + (\nu_j + \omega_j Homework_{ij} + u_{ij}) \end{aligned} \tag{22.16}$$

21 The deviance for the OLS model is 1916.35 and that for the random intercept model is 1842.65, a reduction of about 73.7, which is precisely the LR value in Table 22.7, and this LR value is highly significant.

In Eq. (22.16), the first four terms on the right-hand side are *fixed* and the terms in the parenthesis are random. Model (22.16) is known as a mixed-effects model.

The random effects include the usual regression error term, u_{ij}, the error term v_j associated with the random intercept term, which represents variability among schools, and ω_j representing variability in the slope coefficients across schools.

A noteworthy feature of (22.16) is that it includes an interaction term between *schid* and *Homework*, which brings together variables measured at different levels in hierarchically structured data (Table 22.8).

Table 22.8 Results of Model 22.14: random intercept, random slope, with interaction.

```
. xtmixed math homework cp || schid: homework, variance
xtmixed math homework cp || schid: homework, variance
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log likelihood = –888.11122
Iteration 1: log likelihood = –888.11122
Computing standard errors:
```

Mixed-effects ML regression	Number of obs	=	260
Group variable: schid	Number of groups	=	10
	Obs per group: min	=	20
	avg	=	26.0
	max	=	67
	Wald chi2(2)	=	5.88
Log likelihood = –888.11122	Prob > chi2	=	0.0530

math	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
homework	–.9656993	2.109296	–0.46	0.647	–5.099843	3.168445
cp	.0000805	.000046	1.75	0.080	–9.68e-06	.0001707
_cons	44.82843	2.487166	18.02	0.000	39.95368	49.70319

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
schid: Independent				
var(homework)	13.16657	6.666627	4.880718	35.51905
var(_cons)	55.87579	27.26257	21.47389	145.3907
var(Residual)	43.29007	3.971024	36.16655	51.81667
LR test vs. linear regression:		chi2(2) = 103.91		Prob>chi2=0.0000

Note: LR test is conservative and is provided only for reference.
Note: cp is the cross-product term schid*homework

First, notice that in the **xtmixed** command, after the || sign, we use *schid*, the Level 2 variable, followed by the regressor, *homework*. If we omit this regressor, we will be back to the random intercept but fixed regressor model. Since in this model *homework* is the only regressor, we are assuming that the slope coefficient of this variable varies from school to school.

The results given in this table seem perplexing. The coefficient of homework is negative, although it is statistically insignificant. The coefficient of the cross-product term is positive and is significant at about the 8% level. This suggests that homework combined with school has a positive impact on the test score: Better schools and more homework have positive effect on test scores.

Looking at the random effects parameters in the third part of this table, it seems the variance of the random slope coefficient is significant at the 5% level, suggesting that the slope coefficient is random.

22.8 HLM Mixed models with more than one regressor

Suppose we extend our model with student/teacher ratio (*ratio*) as an additional regressor. With the added regressor, we have several choices. We can let the intercept only be random; or we can let intercept and slope coefficient with respect to homework be random, or we can let intercept and slope coefficient of ratio to be random or we can let the intercept as well as slopes of the two variables be random.

We will consider all these situations, for they point out some interesting differences among these models:

1. HLM with random intercept but fixed slope coefficients of the two regressors

In this model (Table 22.9), both slope coefficients are statistically significant and have the correct signs – math score is positively related to the hours of homework and negatively related to the ratio variable – the higher the student/teacher ratio, the lower the math performance, *ceteris paribus*. Also, note that the variances of both the (random) intercept and the regression error term are statistically significant. From the LR test given in this table, we can say that this model is superior to the OLS model.

2. HLM with random intercept and variable homework coefficient but fixed ratio coefficient

In this model (Table 22.10), the *homework* variable is insignificant, but the *ratio* variable is significant and has the correct sign. The *cp*, the cross product of *homework* and *schid*, is positive and is marginally significant. In other words, the homework coefficient by itself is not significant, but in conjunction with *schid* is it marginally significant. This suggests that *schid* has an attenuating influence on homework.

3. HLM with random intercept and variable ratio coefficient but fixed homework coefficient

In this model (Table 22.11), both slope coefficients are individually statistically highly significant and have the correct signs, but the cross-product between *schid* and *ratio* is not. It may be that the student/teacher ratio does not vary much from school to school.

Table 22.9 HLM with random intercept but fixed slope coefficients.

```
xtmixed math homework ratio || schid:,variance
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log likelihood = –918.19803
Iteration 1: log likelihood = –918.19803
Computing standard errors:
```

Mixed-effects ML regression	Number of obs	=	260
Group variable: schid	Number of groups	=	10
	Obs per group: min	=	20
	avg	=	26.0
	max	=	67
	Wald chi2(2)	=	46.83
Log likelihood = –918.19803	Prob > chi2	=	0.0000

math	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**[95% Conf. Interval]**	
homework	2.218977	.3740691	5.93	0.000	1.485815	2.952139
ratio	–.9342158	.3106066	–3.01	0.003	–1.542993	–.3254381
_cons	59.46378	5.009873	11.87	0.000	49.6446	69.28295

Random-effects Parameters	**Estimate**	**Std. Err.**	**[95% Conf. Interval]**	
schid: Identity				
var(_cons)	10.58601	5.875833	3.566705	31.41935
var(Residual)	64.28677	5.746139	53.95588	76.59572
LR test vs. linear regression:	chibar2(01) = 23.35		Prob >= chibar2 = 0.0000	

4. HLM with random intercept, random slope coefficients, and interaction effects

The only coefficient that is significant in Table 22.12 (at about the 6% level) is the cross-product of homework with school ID. Although this model is better than the OLS model, as judged by the LR statistic, this model points out one of the limitations of HLM if we introduce too many interaction terms, especially if the sample is small, as in the present instance.

Considering the four models, it seems the model with random intercept and random ratio coefficient seems to be the best (Table 22.11). We can use the likelihood ratio test to compare the modes to choose among the four models. We leave this as an exercise for the reader.[22]

We can further extend our model by considering several regressors, such as homework, student/teacher ratio (*ratio*), socio-economic status of students (*ses*), parent education (*parented*), student's sex, and race. The reader is urged to pursue this exercise using the more extended data given in Exercise 22.1.

22 One can also use information criteria, such as the Akaike or Schwarz, that we discussed earlier in the text, to choose among the four models.

Table 22.10 HLM with random intercept, one random coefficient, and one fixed coefficient.

xtmixed math homework ratio cp|| schid: homework, variance
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log likelihood = −886.4015
Iteration 1: log likelihood = −886.4015
Computing standard errors:

Mixed-effects ML regression	Number of obs	=	260
Group variable: schid	Number of groups	=	10
	Obs per group: min	=	20
	avg	=	26.0
	max	=	67
	Wald chi2(3)	=	10.01
Log likelihood = -886.4015	Prob > chi2	=	0.0184

math	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**[95% Conf.**	**Interval]**
homework	−.9382923	2.082309	−0.45	0.652	−5.019544	3.142959
ratio	−1.137399	.5558153	−2.05	0.041	−2.226777	−.0480207
cp	.0000798	.0000454	1.76	0.079	−9.18e−06	.0001688
_cons	62.45449	8.840323	7.06	0.000	45.12778	79.78121

Random-effects Parameters	**Estimate**	**Std. Err.**	**[95% Conf.**	**Interval]**
schid: Independent				
var(homework)	12.82021	6.535745	4.720113	34.82073
var(_cons)	37.15618	19.35477	13.38559	103.1394
var(Residual)	43.37125	3.987358	36.21981	51.93469
LR test vs. linear regression:	chi2(2) = 67.41		Prob > chi2 = 0.0000	

Note: LR test is conservative and provided only for reference.

22.9 Comparison of three approaches[23]

Taking Eq. (22.16) as a prototype of mixed effects models, we have considered three approaches to estimating such models: (1) Classical OLS regression, (2) OLS regression with clustered standard errors, and (3) HLM models.

Classical OLS regression

A critical assumption underlying the classical OLS regression is that the regression errors are $u_{ij} \sim N(0,\sigma^2)$, that is they are NIID. However, this assumption is untenable in clustered data in which the observations are most likely to be correlated. As a result, the standard errors are generally underestimated, as we showed in our illustrative examples.

23 For details, see Primo, D. M., Jacobsmeier, M. L., and Milyo, J. (2007) Estimating the impact of state policies and institutions with mixed-level data, *State Politics and Policy Quarterly*, 7(4), 446–59.

Table 22.11 HLM with random intercept, one random coefficient, and one fixed coefficient.

xtmixed math homework ratio cpr || schid: ratio, variance
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log likelihood = –917.32341
Iteration 1: log likelihood = –917.17633
Iteration 2: log likelihood = –917.13629
Iteration 3: log likelihood = –917.13527
Iteration 4: log likelihood = –917.13527
Computing standard errors:

Mixed-effects ML regression	Number of obs	=	260
Group variable: schid	Number of groups	=	10
	Obs per group: min	=	20
	avg	=	26.0
	max	=	67
	Wald chi2(3)	=	53.58
Log likelihood = –917.13527	Prob > chi2	=	0.0000

math	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**[95% Conf.**	**Interval]**
homework	2.222804	.3721716	5.97	0.000	1.493361	2.952246
ratio	–1.100971	.296897	–3.71	0.000	–1.682878	–.5190637
cpr	3.67e–06	2.38e–06	1.54	0.123	–9.97e–07	8.34e–06
_cons	59.94752	4.516209	13.27	0.000	51.09591	68.79912

Random-effects Parameters	**Estimate**	**Std. Err.**	**[95% Conf.**	**Interval]**
schid: Independent				
var(ratio)	4.56e–15	7.54e–14	3.73e–29	.5567737
var(_cons)	8.037677	4.752642	2.522429	25.61192
var(Residual)	64.28484	5.745975	53.95424	76.59343

LR test vs. linear regression: chi2(2) = 15.77 Prob > chi2 = 0.0004

Note: LR test is conservative and provided only for reference.
Note: *cpr* is cross-product of *schid* and *ratio*

OLS with clustered standard errors

Clustered standard errors of regression coefficients allow for correlation among observations in a cluster, but they assume that the observations across clusters are independent. That is, these standard errors take into account general forms of heteroscedasticity as well as for intra-cluster correlation.[24]

24 For a technical discussion of various types of standard errors, see Angrist, J. S. and Pischke, J̈.-S., *Mostly Harmless Econometrics: An Empiricist's Companion*, Chapter 8. Princeton University Press, Princeton, New Jersey, 2009.

Table 22.12 HLM with random intercept, random slopes, and interaction terms.

xtmixed math homework ratio cp cpr || schid: homework ratio, variance
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log likelihood = –886.00286
Iteration 1: log likelihood = –885.92193
Iteration 2: log likelihood = –885.88619
Iteration 3: log likelihood = –885.88589
Iteration 4: log likelihood = –885.88589
Computing standard errors:

Mixed-effects ML regression	Number of obs	=	260
Group variable: schid	Number of groups	=	10
	Obs per group: min	=	20
	avg	=	26.0
	max	=	67
	Wald chi2(4)	=	11.48
Log likelihood = –885.88589	Prob > chi2	=	0.0216

math	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**[95% Conf.**	**Interval]**
homework	–1.108778	2.08595	–0.53	0.595	–5.197165	2.979608
ratio	–.9298151	.5686153	–1.64	0.102	–2.044281	.1846504
cp	.0000842	.0000456	1.85	0.064	–5.05e–06	.0001735
cpr	–4.71e–06	4.54e–06	–1.04	0.300	–.0000136	4.20e–06
_cons	61.96072	8.458131	7.33	0.000	45.38309	78.53835

Random-effects Parameters	**Estimate**	**Std. Err.**	**[95% Conf.**	**Interval]**
schid: Independent				
var(homework)	12.77393	6.508705	4.705563	34.67668
var(ratio)	1.87e–12	.	.	.
var(_cons)	33.45384	17.46507	12.0244	93.074
var(Residual)	43.35486	3.983863	36.20939	51.91041

LR test vs. linear regression: chi2(3) = 68.21 Prob > chi2 = 0.0000
Note: LR test is conservative and is provided only for reference.

Hierarchical linear models (HLM)

In HLM we explicitly model the compound error term, as in Eq. (22.16). That is, it allows us to estimate how much Level 1 and Level 2 (or higher levels) variables contribute to the overall error term. It also enables us to estimate the individual error variances and their covariances. The one big difference between clustered standard errors and HLM is that in the former the denominator degrees of freedom is based on the number of observations, whereas in the latter it is the number of clusters. Since in implementing HLM the variability in the regressors and the residuals is taken into account, the point estimates of the regression coefficients are different from those of OLS with robust standard errors or with clustered standard errors, and so are the standard errors of the estimators.

Since HLMs are estimated by maximum-likelihood estimation (MLE), if there is misspecification of the compound error term (such as Eq. (22.16)), it propagates throughout the HLM estimation procedure, including the estimation of regression coefficients.[25] Thus it is critically impotant to get the HLM specification correct for point estimation as well as for statistical inference. In contrast, since the clustered standard errors are calculated after estimation, this approach does not have the risks associated with HLM. Since HLM is data- and computation-intensive, it usually does not work well if there are too few clusters. Also, if there are many observations and many cross-level interactions, the analysis may become unduly complex or unwieldy.

Therefore, the choice between clustered standard errors OLS models and multilevel models is not always easy in practice. An important factor for the practitioner is the number of clusters. If you have relatively few clusters, multilevel modeling may not be appropriate, for you need to have a fair number of clusters to implement HLM. That is why Steenbergen and Jones, who are proponents of HLM, advise that since HLM make heavy demands on theory and data, such models should not be used blindly in analyzing multilevel data.[26]

22.10 Some technical aspects of HLM

As noted earlier, HLM models are estimated by the method of maximum-likelihood (ML). There are two variants of ML: Full Maximum Likelihood (FML) and Restricted Maximum Likelihood (RML). In FML, the regression coefficients and the variance components are included in the likelihood function (LF). In RLM only the variance components are included in the LF. As Hox notes,[27]

> the difference is that FML treats the estimates for the regression coefficients as known quantities when the variance components are estimated, while RML treats them as estimates that carry some amount of uncertainly. Since RML is more realistic, it should in theory, lead to better estimates when the number of groups [clusters] are small. FML has two advantages over RML; the computations are generally easier, and since the regression coefficients are included in the likelihood function, the likelihood ratio can be used to test for differences between two nested models that differ only in the fixed part (the regression coefficient). With RML only differences in the random part (the variance components) can be tested this way.

In HLM, the Wald test is used for testing statistical significance, but it assumes that we have fairly a large number of groups or equal group sizes. The power of the Wald test for testing the significance of individual regression coefficients depends on the total number of observations in the sample. But the power of the tests for higher level effects (Level 2, Level 3, etc.) and cross-level interaction effects, depends more on the number of groups than the number of observations.

---

25 But remember that misspecification of the error term is bad for all estimation methods.

26 Steenbergen, M. R. and Jones, B. S. (2002) Modeling multilevel data structures, *American Journal of Political Science*, **46**(1), 218–37.

27 Hox, *op cit*., pp. 147–54.

The simulation experiments done by various authors suggest a trade-off between sample sizes at different levels of hierarchy.[28] The literature suggests that increasing the sample sizes at all levels of hierarchy increases the accuracy of the regression coefficients as well as their standard errors. Kreft has proposed the 30/30 rule – a sample of at least 30 groups and at least 30 individuals per group.[29]

If interest is on cross-level interactions effects, Hox suggest the 50/20 rule – 50 groups with about 20 individuals per group. But if one is interested in the random effects part of HLM, including variance and covariance components, he suggests the 100/10 rule – 100 groups with about 10 individuals per group. Of course, in using these rules of thumb, one should not forget the costs involved in collecting and analyzing data with more groups and or more observations.

22.11 Summary and conclusions

The primary goal of this chapter was to introduce the reader to the rapidly evolving field of hierarchical linear models (HLM). HLM has useful applications in a variety of fields, as mentioned in the introduction.

In analyzing hierarchical or multi-level data, we have three choices: OLS with robust standard errors, OLS with clustered standard errors, and multilevel modeling. The standard OLS model assumes, among other things, that the regression errors are identically and independently distributed as a normal distribution with zero mean and constant variance. In hierarchical data, such assumptions is not usually tenable because observations within a cluster or group (say, students in a class) are likely to be correlated due to environmental factors.

OLS with clustered standard errors is an improvement over the standard OLS method because it at least takes into account correlation within a cluster. But it assumes that the errors among clusters are uncorrelated. The point estimates of the regression parameters are identical to those obtained from OLS with robust standard errors, but the clustered standard errors are generally higher than those of the standard OLS model.

Since neither the standard OLS nor OLS with cluster standard errors approach takes into account intra-class correlations (ICC) nor the correlation among clusters, an alternative is the HLM. In HLM, we take these factors into account and it also provides estimates of the variances and covariances of the component error terms. The HLM coefficients and their standard errors are generally different from the other two approaches.

We also considered some technical aspects of HLM, such as the difference between Full Information Maximum Likelihood (FML) and Restricted Maximum Likelihood (RML) estimation of HLMs. In addition, we also discussed the appropriate number of individual and group observations to carry out several aspects of HLM.

In this chapter we have considered only the linear hierarchical models. But there are several multilevel nonlinear models that can be estimated in *Stata* by using the commands **xtlogit**, **xtprobit**, **xttobit**, **xtpoisson**, **xtmelogit** and **xtmepoisson**. The

---

28 See, Mok, M., *Sample Size Requirements for 2-level Designs in Educational Research*. Multilevel Models Project, University of London, 2005.

29 See Kreft, I. G. G., *Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies*, California State University, Los Angeles, 1996.

interested reader may consult the *Stata* manuals (or SAS or SPSS manuals) for further details.

Exercises

22.1 In this chapter we discussed HLM modeling of math test data for 260 students in 10 randomly selected school. **Table 22.1** (on the companion website)[30] gives data on 519 students in 23 schools – 8 schools are in the private sector and 15 schools are in the public sector. The student level (Level 1) data and the school level data (Level 2) are the same as in the sample discussed in the text.

Explore these data by developing HLM model(s), considering the relevant explanatory variables and taking into account various cross-level interaction effects and compare your analysis with the standard OLS regression using clustered standard errors.

22.2 There are many interesting data sets given in Sophia Rabe-Hesketh and Andres Skrondal's *Multilevel and Longitudinal Modeling Using Stata*, Vol. 1 (continuous response models) and Vol. 2 (categorical responses, counts and survival), 3rd edn, published by Stata Press. All the data in these volumes can be downloaded from the following website:

http://www.stata-press.com/data/mlmus3.html

Choose the data of your interest and try to model it using HLM, considering various aspects of HLM modeling.

30 The data were adapted from Kreft, I. and De Leeuw, J. *Introducing Multilevel Modeling*, Sage Publications, California, 2007.